PharmaSUG 2012 - Paper AD22

Using JMP® Partition to Grow Decision Trees in Base SAS®

Mira Shapiro, Analytic Designers LLC, Bethesda, MD

ABSTRACT

Decision Tree is a popular technique used in data mining and is often used to pare down to a subset of variables for more complex modeling efforts. If your organization has only licensed Base SAS and SAS/STAT you may be surprised to find that there is no procedure for decision trees. However, if you are licensed JMP 9 user, you can build and test a decision tree with JMP. The Modeling→Partition analysis provides an option for creating SAS Data Step scoring code. Once created, the scoring code can be run in Base SAS. This discussion with provide a brief overview of decision trees and illustrate how to create a decision tree with Partition in JMP and then create the SAS Data Step Scoring code.

INTRODUCTION

As analysts and statisticians we are often faced with the question of what relationships are present in our data. JMP provides methods to find an answer. The JMP Partition Platform provides an easy way to create decision trees with numerous options for perfecting and interpreting the results. This discussion will focus on creating a decision tree and will give an introduction to the many options that are available for improving and interpreting the results within JMP. Additionally, the option to create a decision tree in JMP and export the code to be run in Base SAS will be demonstrated.

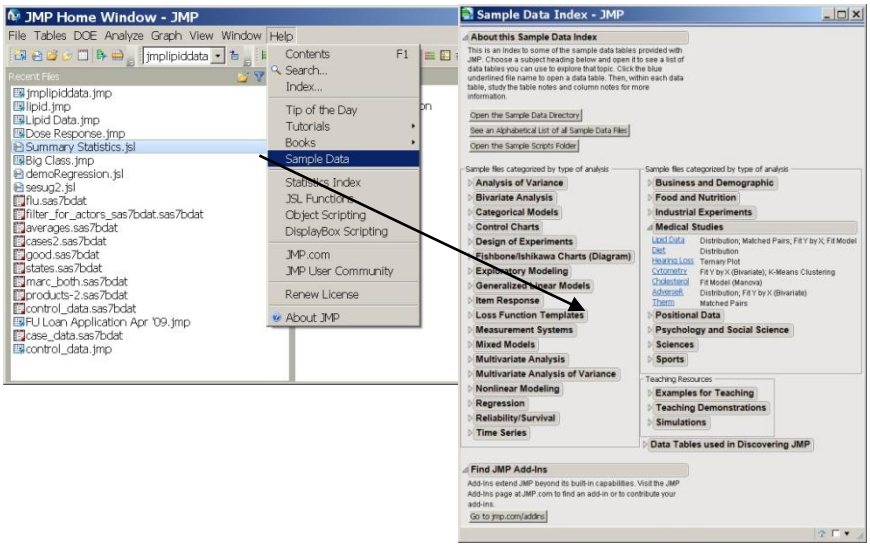
WHAT IS A DECISION TREE?

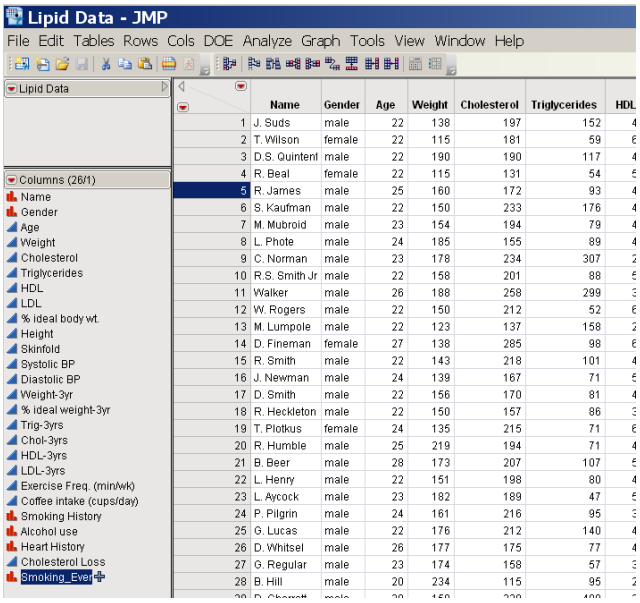
A decision tree model allows for examination of the relationship between a response variable and multiple possible predictors. The potential predictors are evaluated using statistical methods appropriate to their type and assessed as to their predictive value for the response variable. The data is then split into two groups based on the value of the predictor. As the tree is built by recursive splitting, the predictors are re-evaluated at each stage.

There are numerous resources available for understanding the underlying statistical techniques and algorithms used in the decision tree modeling process. Such a discussion is beyond the scope of this paper. Several recommended books, papers and online resources are listed in the References and Recommended Reading sections at the end of this paper.

JMP SAMPLE DATA USED IN EXAMPLES

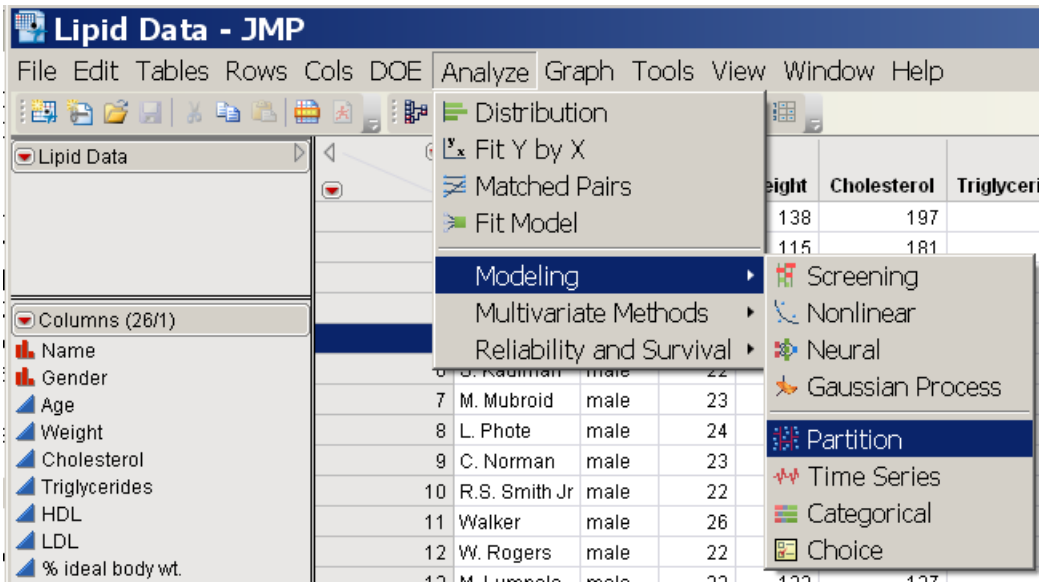
A variety of sample data sets are included with JMP. The index of data sets is helpful to find a sample data set by subject area or analysis type and provides opportunities for new users to explore JMP's capabilities. The examples in this paper use the sample data set "Lipid Data". The sample JMP data sets are found under the *JMP Home Window→ Help→Sample Data***.**



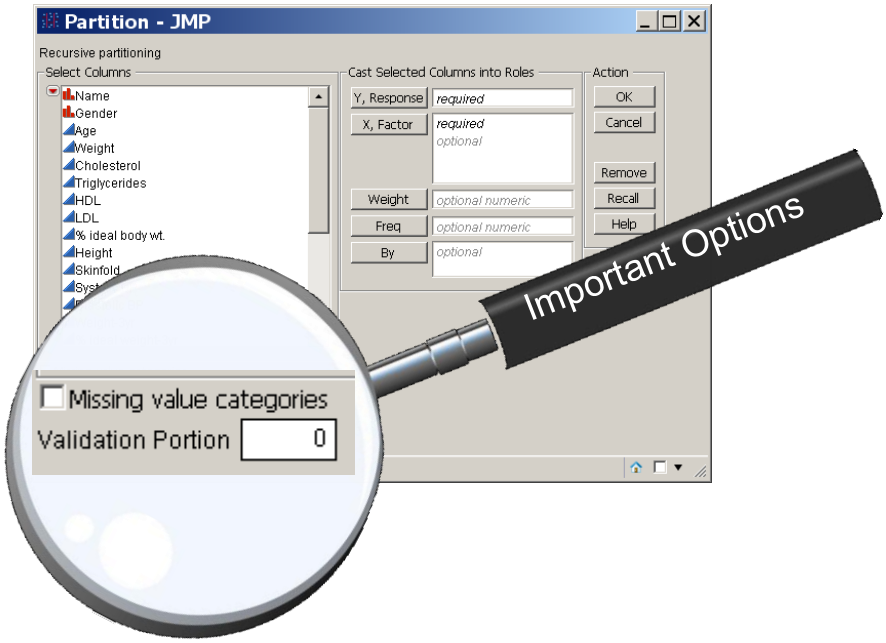


GROWING A DECISION TREE

Once you have opened or created your data table in JMP, there are multiple ways to invoke the JMP Partition Platform. From the JMP Home Window, select ***Analyze → Modeling → Partition*** to begin.



Next, the JMP Partition selection pane opens, providing the opportunity to select variables for roles in the model and make use of the initial options for the model.



Missing Value Categories:

Y Response

	Checked	Unchecked
Categorical	Additional variable level is created for the missing values	Excluded
Continuous	Excluded	Excluded

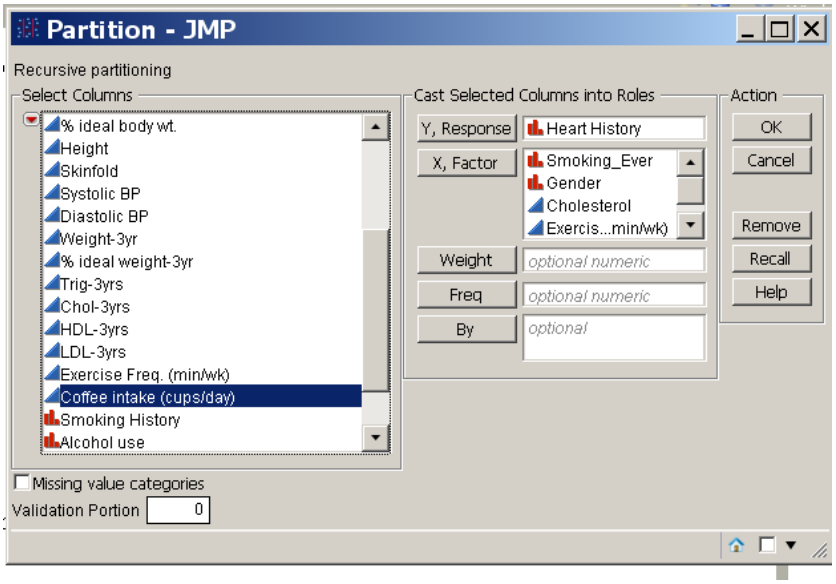
X Predictors (used as splitting variable)

	Checked	Unchecked
Categorical	Additional variable level is created for the missing values	Random assignment of value to one side of the split
Continuous	Random assignment of value to one side of the split	Random assignment of value to one side of the split

Validation Portion:

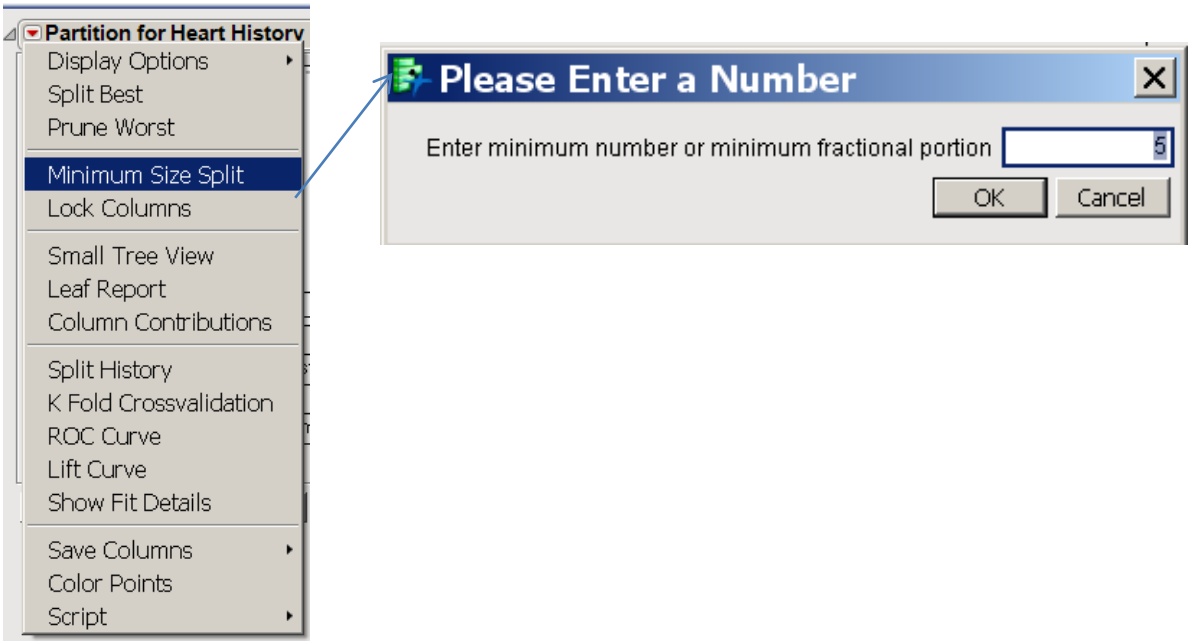
This option allows the user to use a portion of their data to estimate the model parameters, leaving the remaining portion to validate the model.

At this point, drag and drop columns to cast them into roles. For this example, *Heart History* is chosen as the Y response, and several others are cast into the role as potential predictors.

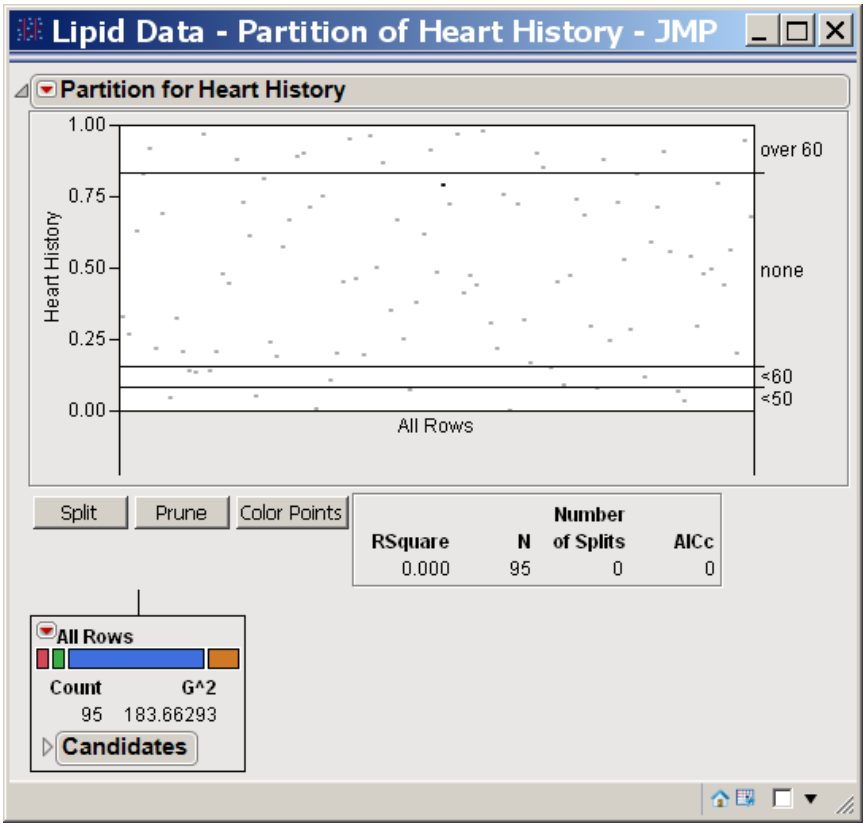


Minimum Size Split

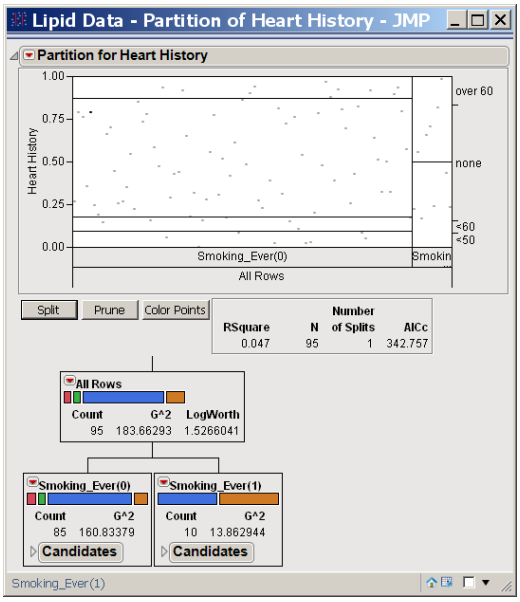
The user has the option to define the minimum size of a group in a split. To access this option, click the hotspot next to the main partition window, and mouse down to ***Minimum Size Split.*** The size can be entered as a number or a fractional portion. The Partition process will not create a group that violates this criterion.



Once the ***OK*** button is clicked, the Partition Platform is launched. The points are shown on the initial pane, with the values of ***Heart History*** shown on the right. It is now up to the user to click the ***Split*** button to begin the process.

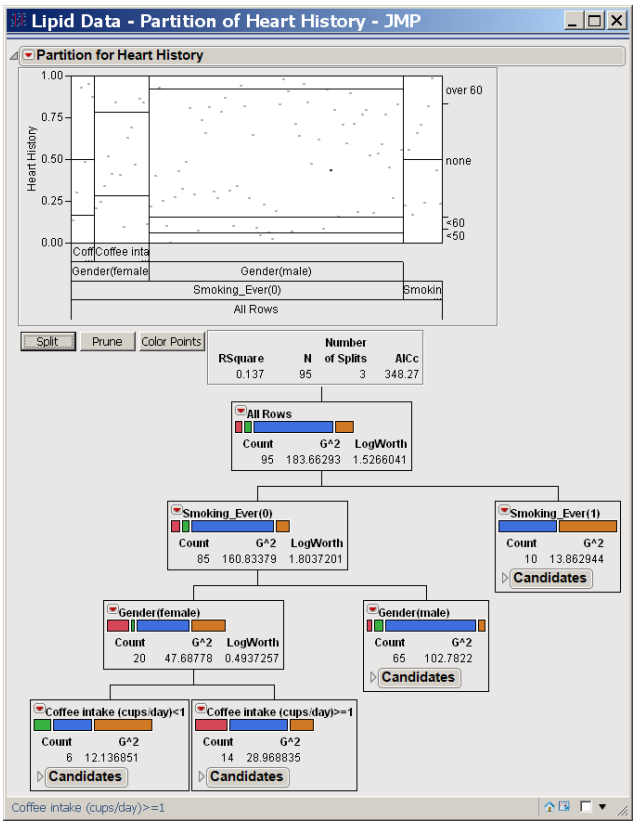


After the ***Split*** button is clicked below the main Partition window, JMP begins to create the decision tree. The first split selected is the ***Smoking_Ever*** variable that was created from the smoking history variable and coded as 0/1.



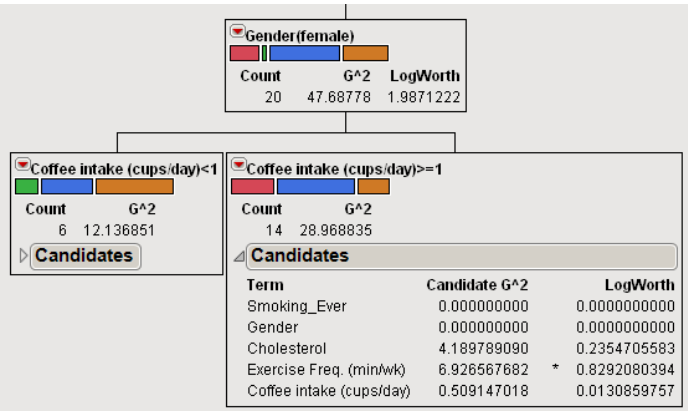
SPLITTING AND PRUNING YOUR DECISION TREE

Partition is a recursive process, and the user has multiple options for continuing and controlling the process. The ***Split*** button was clicked again; the "best" split selected by JMP was ***Gender*** for non-smokers (***Smoking_Ever***=0), and the next was ***Coffee Intake*** for females.

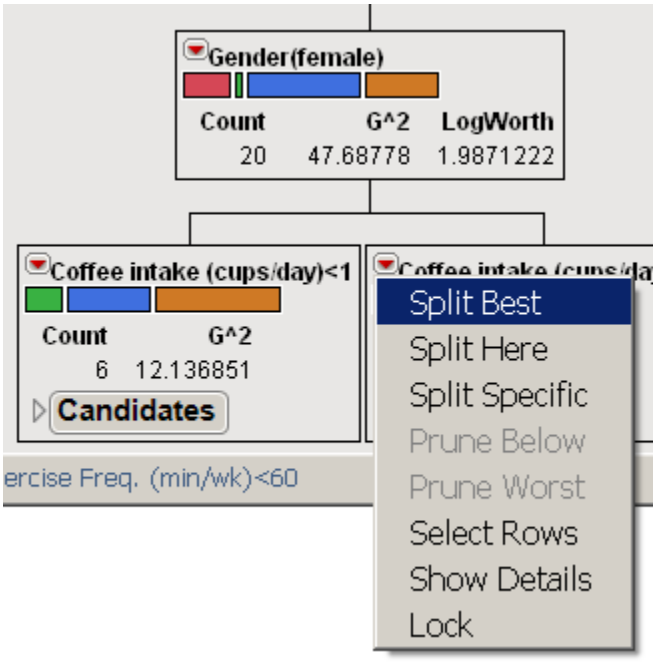


Picking a Candidate

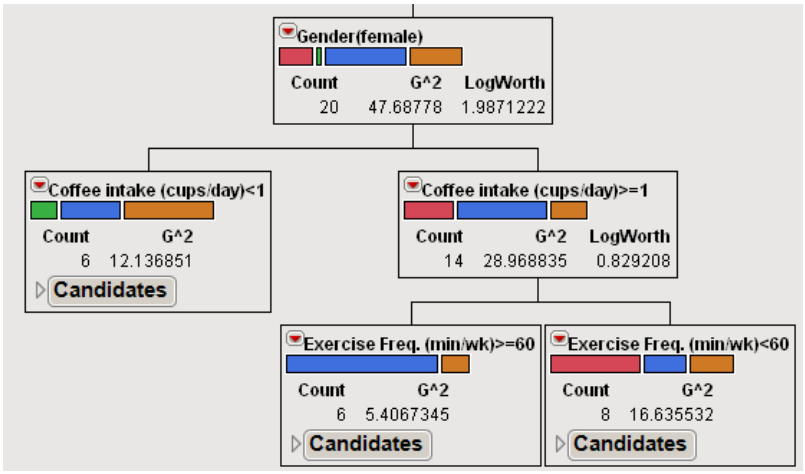
In many cases the user has some subject area knowledge that may direct them to pick a particular split that may not show up as the "best" in the JMP evaluation criteria. By clicking the gray triangle to the left of ***Candidates*** in any node, the user can examine all of the variables that can be used for a split and their associated statistics.



Alternatively, clicking under the hotspot at any node will reveal available options. Here we click ***Split Best*** to allow JMP to select the variable for the split.

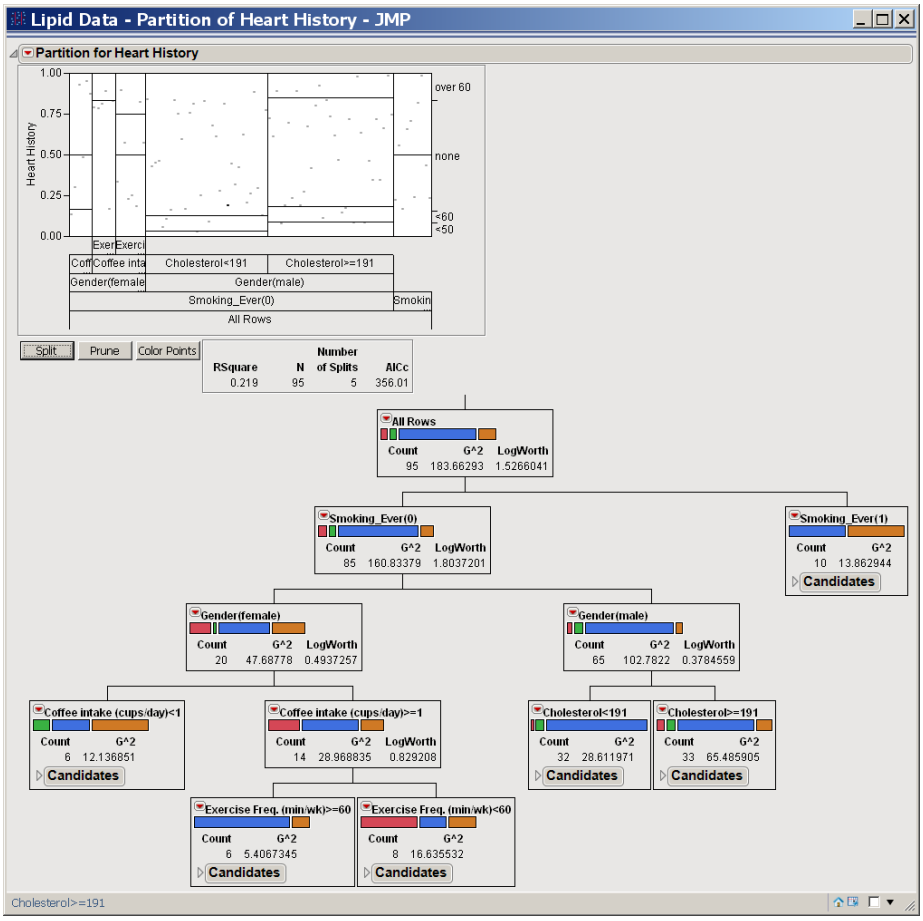


The best split under ***Coffee Intake*** of 1 cup a coffee or greater per day is ***Exercise Freq.***



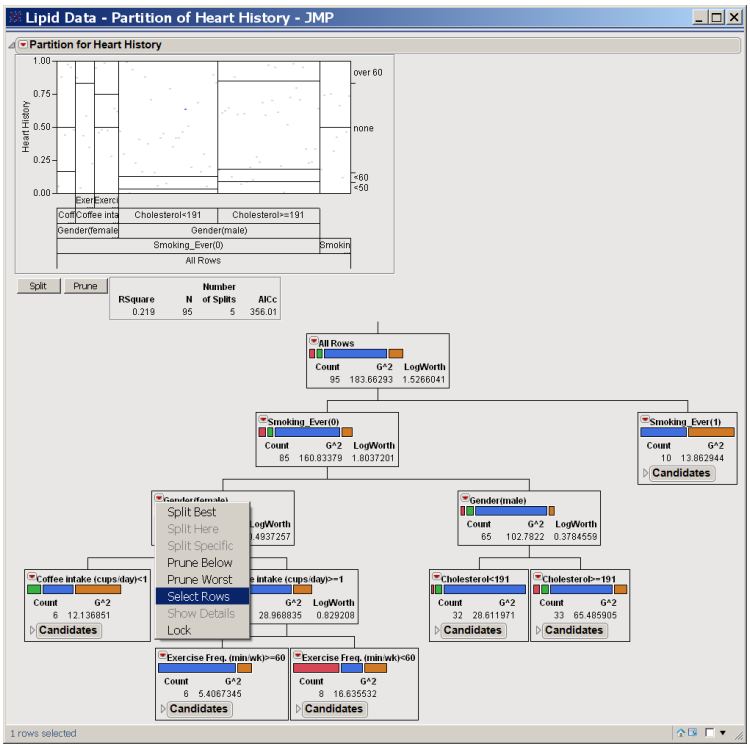
Recursive Splitting

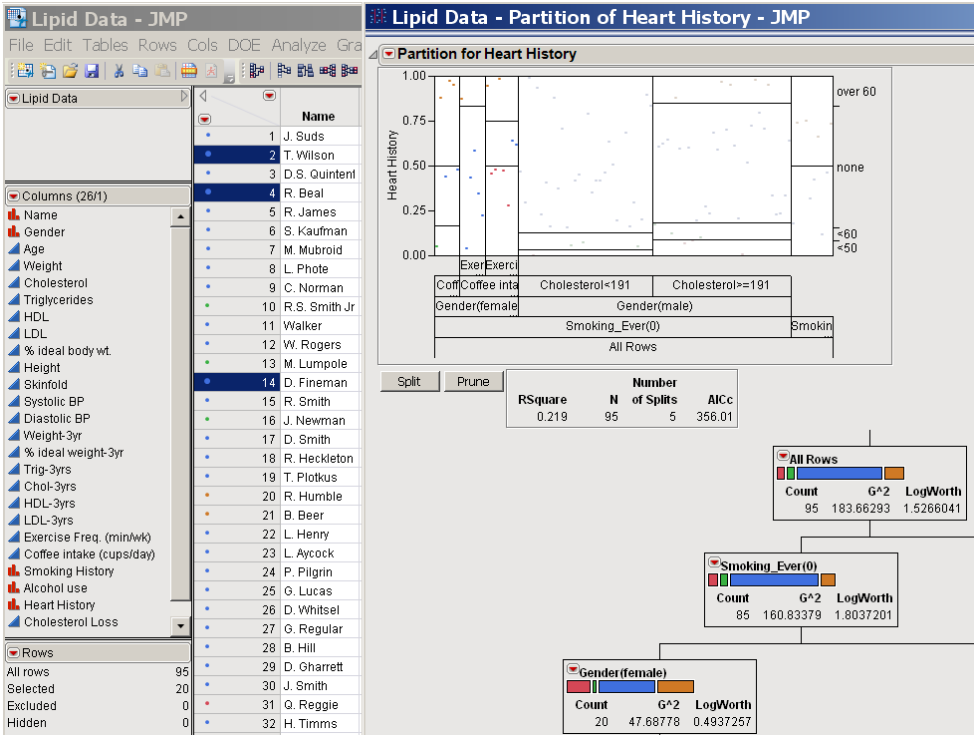
The user can use a combination of techniques to build their decision tree model. No node will be created that violates the minimum split criterion specified in the Partition Platform Launch Window.



Dynamic Linking

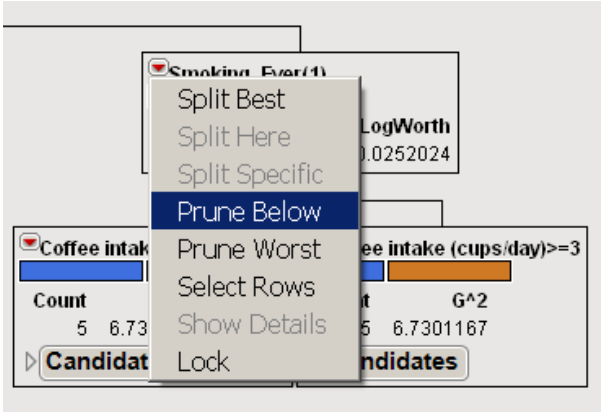
As in all other JMP operations, Dynamic Linking of the data table and results allows the user to visualize the relationships. Clicking on the hotspot on any node and selecting ***Select Rows*** causes the data table to become visible, and the associated rows will be highlighted.

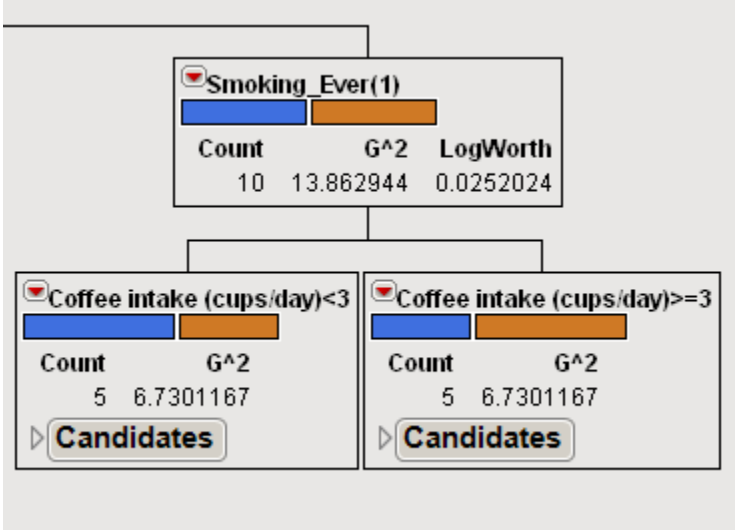


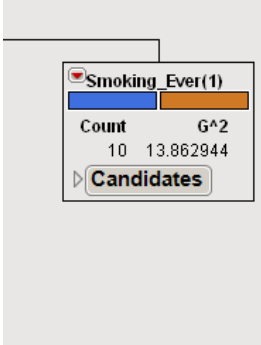


Pruning A Decision Tree

Just as the tree can be built with JMP, nodes can be removed by pruning areas of the tree. Looking at the level that we added in the previous step, we've decided to remove the ***Coffee Intake*** split. By clicking the hot spot next to ***Smoking_Ever,*** we reveal the available options, including ***Prune Below.*** Selecting ***Prune Below*** will result in the removal of all notes below the selected node.

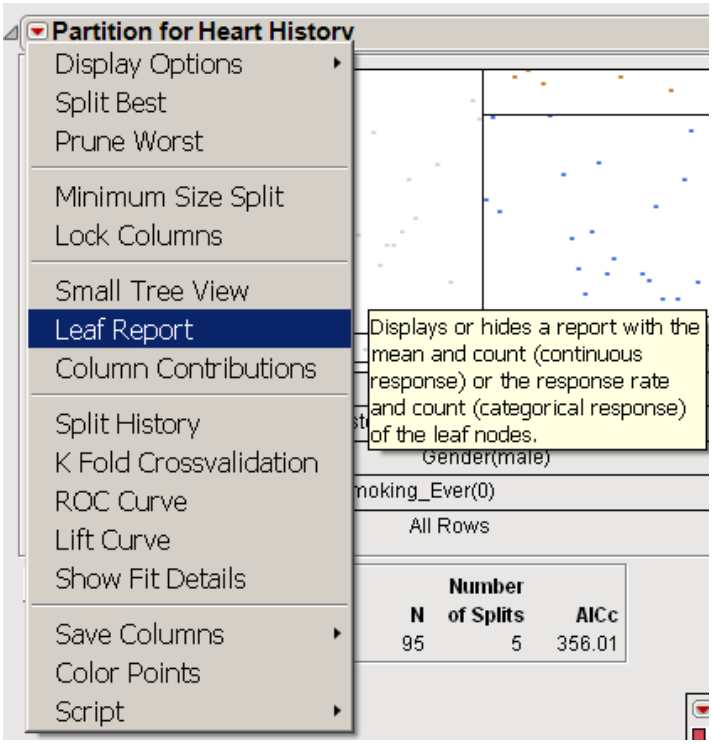




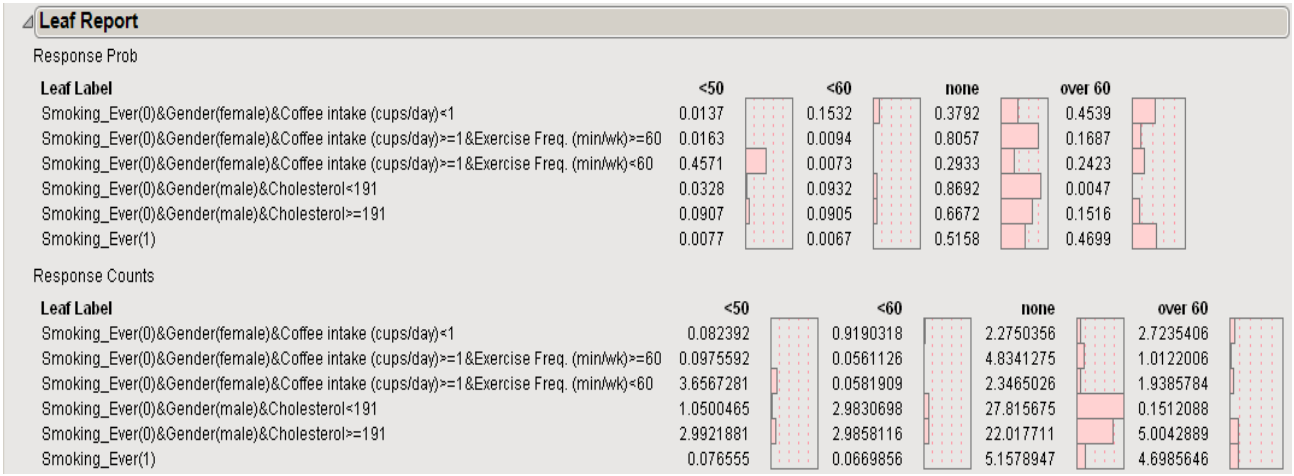


OPTIONS FOR UNDERSTANDING AND EVALUATING YOUR DECISION TREE

There are numerous statistics and reports that are available for evaluating your decision tree model. Generating and showing or hiding each of these options is accomplished by clicking the hotspots and making use of the hide/unhide icon. As shown in the following examples, most of the options can be moused over to reveal a box with a description of the particular option. The JMP Help facility and online resources have detailed information about the statistics generated.



Leaf Report



Leaf Report

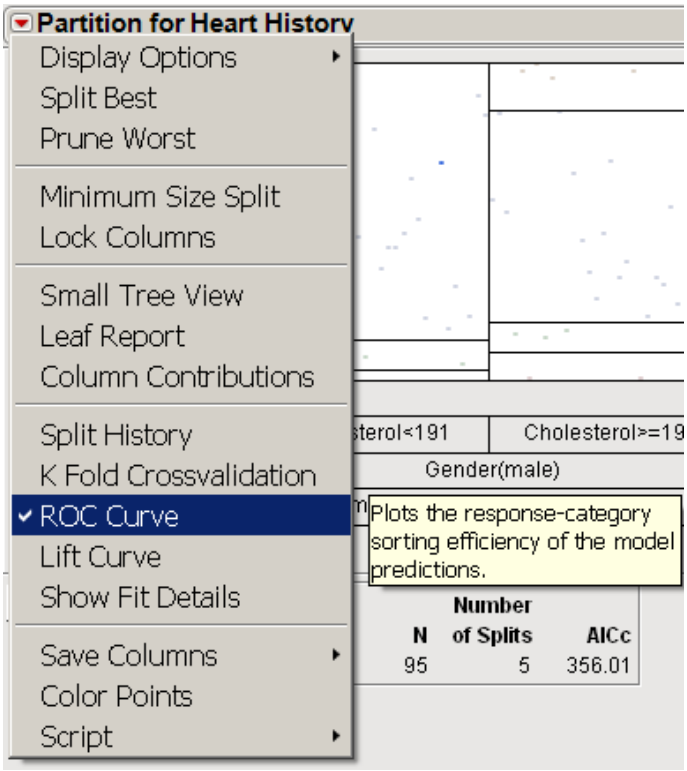
Response Prob

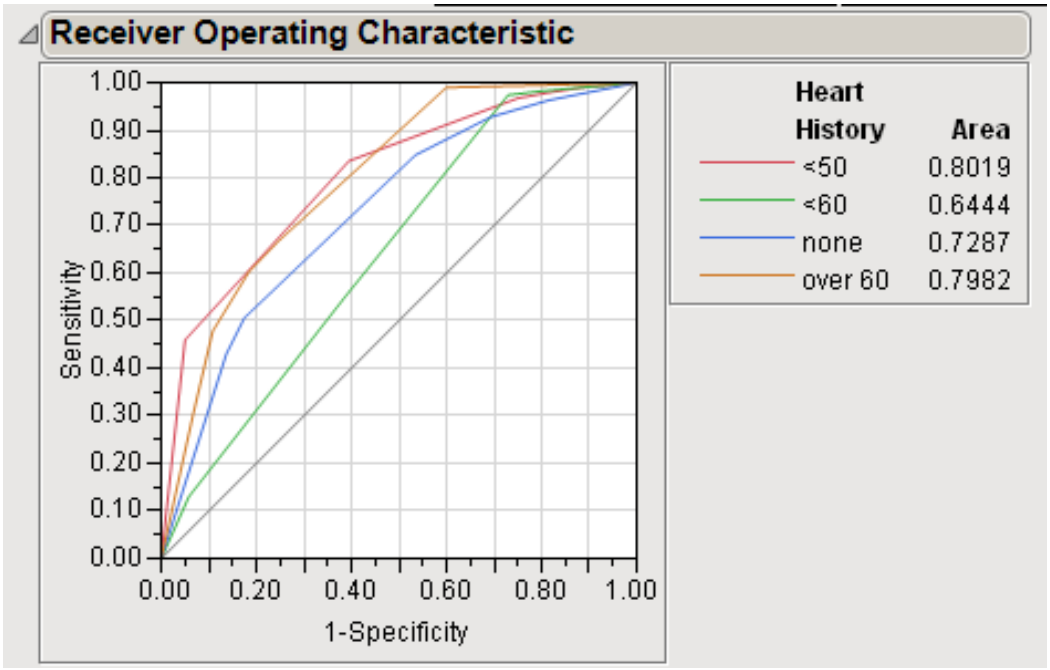
Leaf Label	<50	<60	none	over 60
Smoking_Ever(0)&Gender(female)&Coffee intake (cups/day)<1	0.0137	0.1532	0.3792	0.4539
Smoking_Ever(0)&Gender(female)&Coffee intake (cups/day)>=1&Exercise Freq. (min/wk)>=60	0.0163	0.0094	0.8057	0.1687
Smoking_Ever(0)&Gender(female)&Coffee intake (cups/day)>=1&Exercise Freq. (min/wk)<60	0.4571	0.0073	0.2933	0.2423
Smoking_Ever(0)&Gender(male)&Cholesterol<191	0.0328	0.0932	0.8692	0.0047
Smoking_Ever(0)&Gender(male)&Cholesterol>=191	0.0907	0.0905	0.6672	0.1516
Smoking_Ever(1)	0.0077	0.0067	0.5158	0.4699

Response Counts

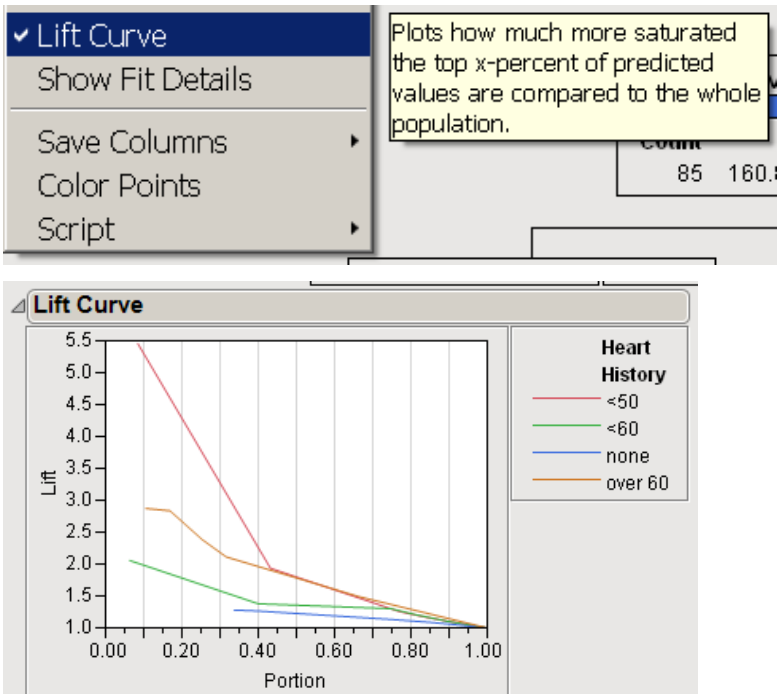
Leaf Label	<50	<60	none	over 60
Smoking_Ever(0)&Gender(female)&Coffee intake (cups/day)<1	0.082392	0.9190318	2.2750356	2.7235406
Smoking_Ever(0)&Gender(female)&Coffee intake (cups/day)>=1&Exercise Freq. (min/wk)>=60	0.0975592	0.0561126	4.8341275	1.0122006
Smoking_Ever(0)&Gender(female)&Coffee intake (cups/day)>=1&Exercise Freq. (min/wk)<60	3.6567281	0.0581909	2.3465026	1.9385784
Smoking_Ever(0)&Gender(male)&Cholesterol<191	1.0500465	2.9830698	27.815675	0.1512088
Smoking_Ever(0)&Gender(male)&Cholesterol>=191	2.9921881	2.9858116	22.017711	5.0042889
Smoking_Ever(1)	0.076555	0.0669856	5.1578947	4.6985646

ROC Curve

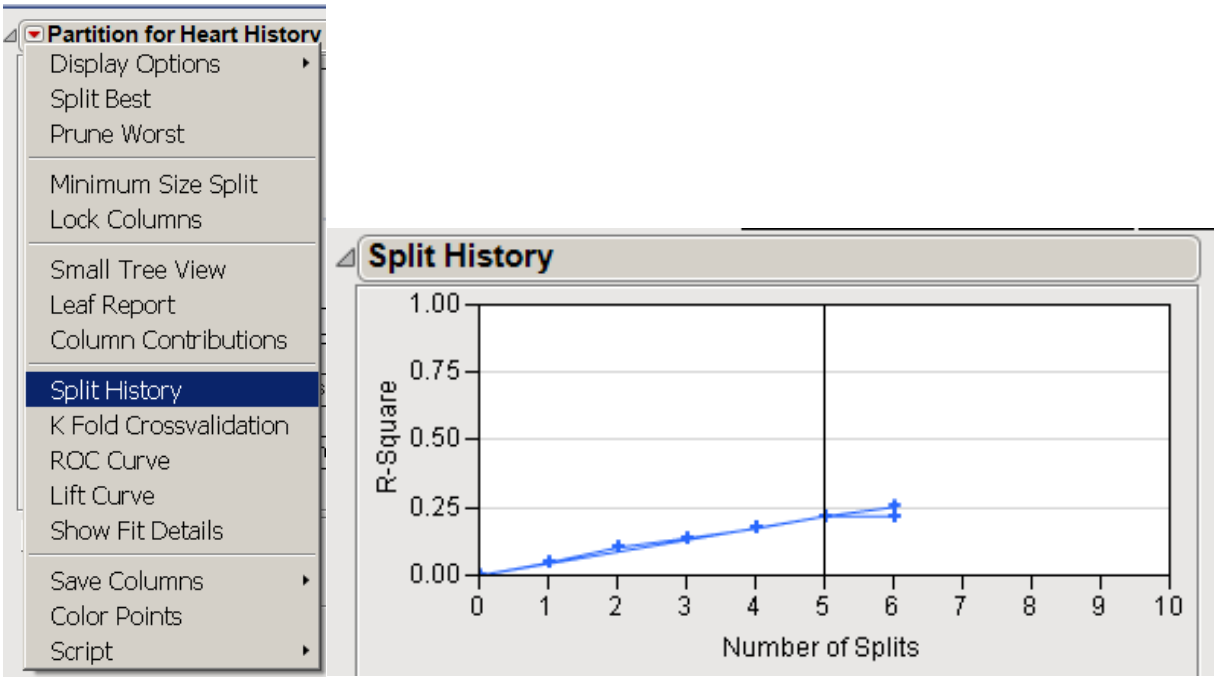

Partition for Heart History
Display Options
Split Best
Prune Worst
Minimum Size Split
Lock Columns
Small Tree View
Leaf Report
Column Contributions
Split History
K Fold Crossvalidation
ROC Curve
Lift Curve
Show Fit Details
Save Columns
Color Points
Script
sterol<191
Cholesterol>=19
Gender(male)
Plots the response-category sorting efficiency of the model predictions.
N
Number of Splits
AICc
95
5
356.01


Receiver Operating Characteristic
Sensitivity
1-Specificity
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00
Heart History
Area
<50
0.8019
<60
0.6444
none
0.7287
over 60
0.7982

Lift Curve



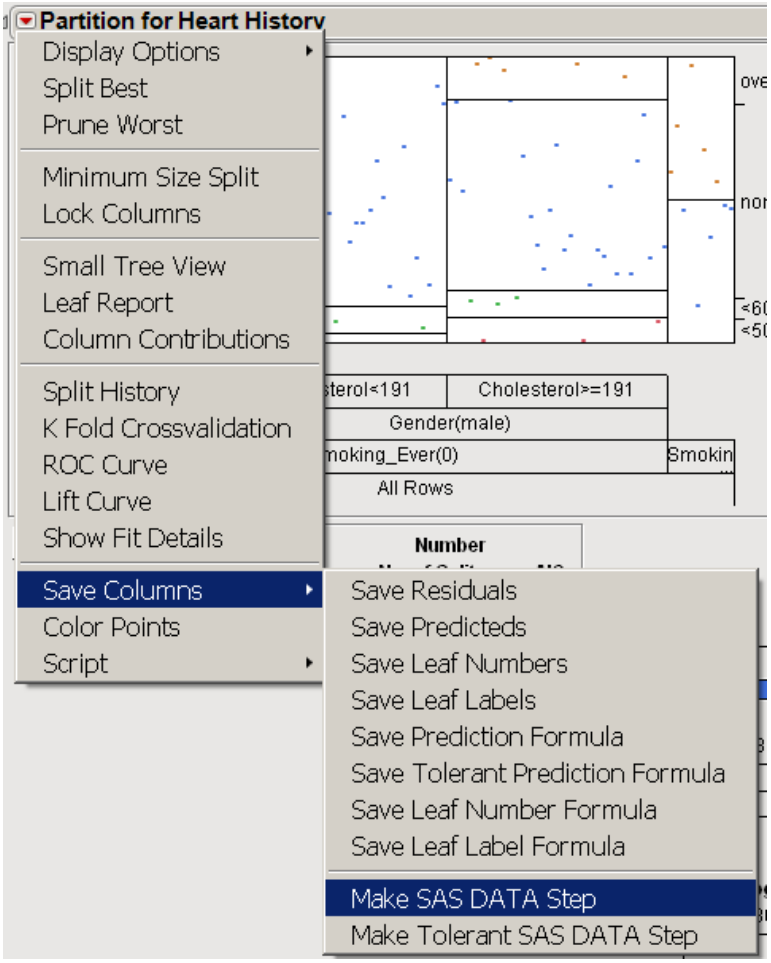
Split History



TRANSPLANTING YOUR DECISION TREE IN SAS

There are times when a user wants to build their decision tree model in JMP and then use the model in Base SAS. In order to create the SAS Data Step code, click on the hotspot at the left of the main partition window, and mouse down to ***Make SAS Data Step***. As shown below, the code will appear in a window entitled ***Partition SAS Scoring Code-JMP***. This code can be run within base SAS with suitable data sources.

Note that the variable names are created using SAS naming conventions, thereby avoiding the need to surround a variable name with special characters with quotes and following it with an 'n'. For example, ***Heart History*** becomes ***Heart_History***. If you are using another data set, it will be necessary to adjust either the program code or your variables' names accordingly.



Partition SAS Scoring Code - JMP

File Edit Tables DOE Analyze Graph Tools View Window Help

```
/*!PRODUCER: JMP - Partition - Decision Tree */
/*!DATA: Lipid Data */
/*!TARGET: Heart_History */
/*!OUTPUT: Prob__50, Heart_History, "<50" */
/*!OUTPUT: Prob__60, Heart_History, "<60" */
/*!OUTPUT: Prob_none, Heart_History, "none" */
/*!OUTPUT: Prob_over_60, Heart_History, "over 60" */
/*!INPUT: Smoking_Ever */
/*!INPUT: Gender */
/*!INPUT: Cholesterol */
/*!INPUT: Exercise_Freq_min_wk_ */
/*!INPUT: Coffee_intake_cups_day_ */


LABEL Prob__50= 'Predicted: Heart_History=<50';
LABEL Prob__60= 'Predicted: Heart_History=<60';
LABEL Prob_none= 'Predicted: Heart_History=none';
LABEL Prob_over_60= 'Predicted: Heart_History=over 60';

Prob__50=0;
Prob__60=0;
Prob_none=0;
Prob_over_60=0;
IF Smoking_Ever=0 THEN DO;
 IF Gender='female'  THEN DO;
  IF Coffee_intake_cups_day_<1 THEN DO;
                                   Prob__50=Prob__50+0.0137320044296789;
                                   Prob__60=Prob__60+0.153171966461003;
                                   Prob_none=Prob_none+0.379172599272267;
                                   Prob_over_60=Prob_over_60+0.453923429837051;
  END;
  ELSE DO;
   IF Exercise_Freq_min_wk_>=60 THEN DO;
                                   Prob__50=Prob__50+0.0162598744924326;
                                   Prob__60=Prob__60+0.00935210673416653;
                                   Prob_none=Prob_none+0.805687918578284;
                                   Prob_over_60=Prob_over_60+0.168700100195117;
   END;
   ELSE DO; Prob__50=Prob__50+0.457091013494114;
            Prob__60=Prob__60+0.00727386079324064;
            Prob_none=Prob_none+0.293312825560888;
            Prob_over_60=Prob_over_60+0.242322300151758;
   END;
  END;
 END;
 ELSE DO;
  IF Cholesterol<191 THEN DO; Prob__50=Prob__50+0.0328139523644141;
                              Prob__60=Prob__60+0.0932209319184951;
                              Prob_none=Prob_none+0.869239841949033;
                              Prob_over_60=Prob_over_60+0.00472527376805779;
  END;
  ELSE DO; Prob__50=Prob__50+0.0906723655301667;
           Prob__60=Prob__60+0.0904791398032452;
           Prob_none=Prob_none+0.667203376009356;
           Prob_over_60=Prob_over_60+0.151645118657233;
  END;
 END;
END;
ELSE DO; Prob__50=Prob__50+0.0076555023923445;
         Prob__60=Prob__60+0.00669856459330144;
         Prob_none=Prob_none+0.515789473684211;
         Prob_over_60=Prob_over_60+0.469856459330144;
END;
```

CONCLUSION

This discussion has provided an overview to the JMP Partition Platform. This facility provides an approach to creating decision trees that is interactive and does not require programming skills. The user has the ability to control the split size, variables and decision tree shape through multiple options within JMP as well as to export Data Step code to SAS. A comprehensive set of statistics and results are available to evaluate the resulting JMP decision tree model.

REFERENCES

http://support.sas.com/resources/papers/proceedings09/TOC.html.

Gaudard, Martha, Ph.D. et. al., "Interactive Data Mining and Design of Experiments: the JMP® Partition and Custom Design Platforms". March 2006, Available at http://www.jmp.com/software/whitepapers/pdfs/372455_interactive_datamining.pdf.

http://www.jmp.com/software/whitepapers/partition_platform/index.shtml

HTTP://JMP.COM/SUPPORT/

THE JMP 9.02 HELP FACILITY

http://www.jmp.com/applications/data_mining/

MASTERING JMP®: AN OVERVIEW OF DATA EXPLORATION AND THE JMP® SCRIPTING LANGUAGE, COURSE NOTES, SAS INSTITUTE INC.

RECOMMENDED READING

Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, Third Edition, by Michael Berry and Gordon Linoff, John Wiley Sons, Inc., April 2011

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

Mira Shapiro
Analytic Designers LLC, Bethesda, MD
mira.shapiro@analyticdesigners.com