

## How to replicate cross-validation in the decision tree

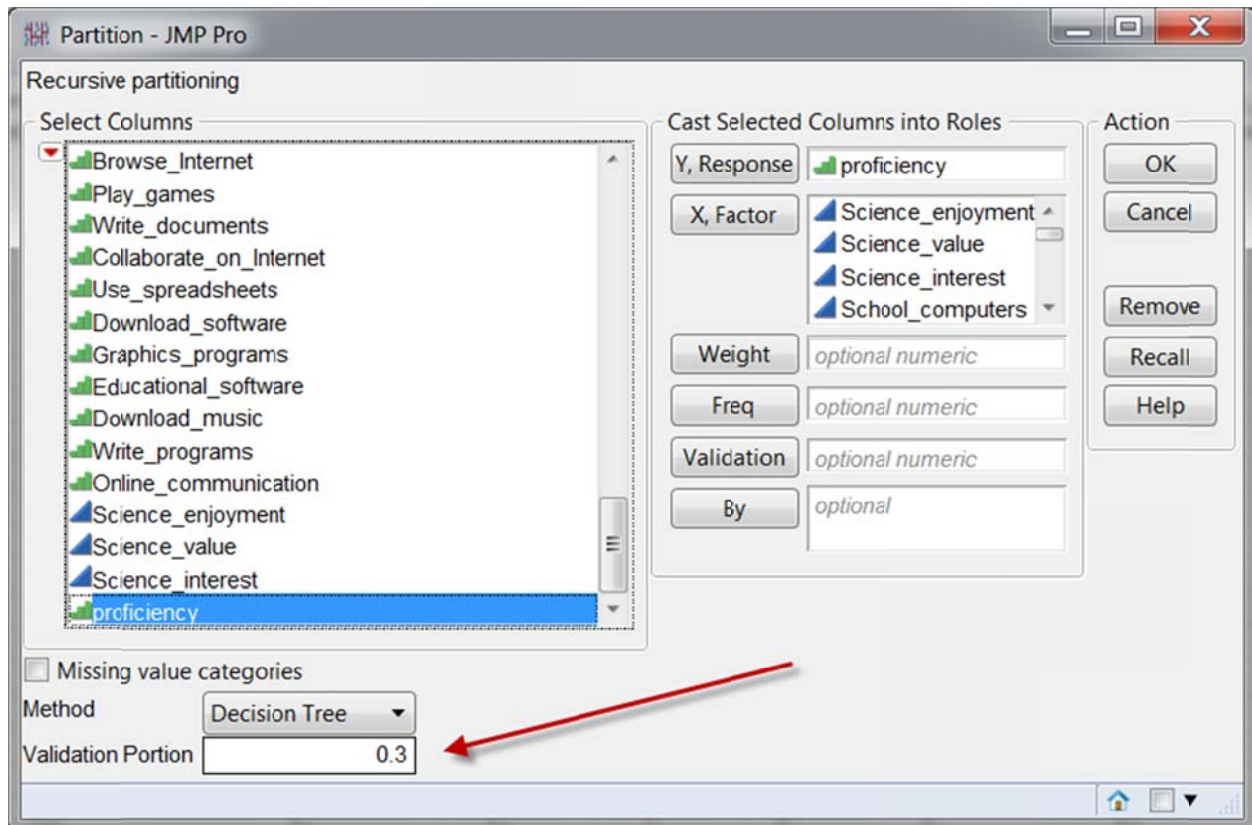
Chong Ho Yu, Ph.D.s (2013)

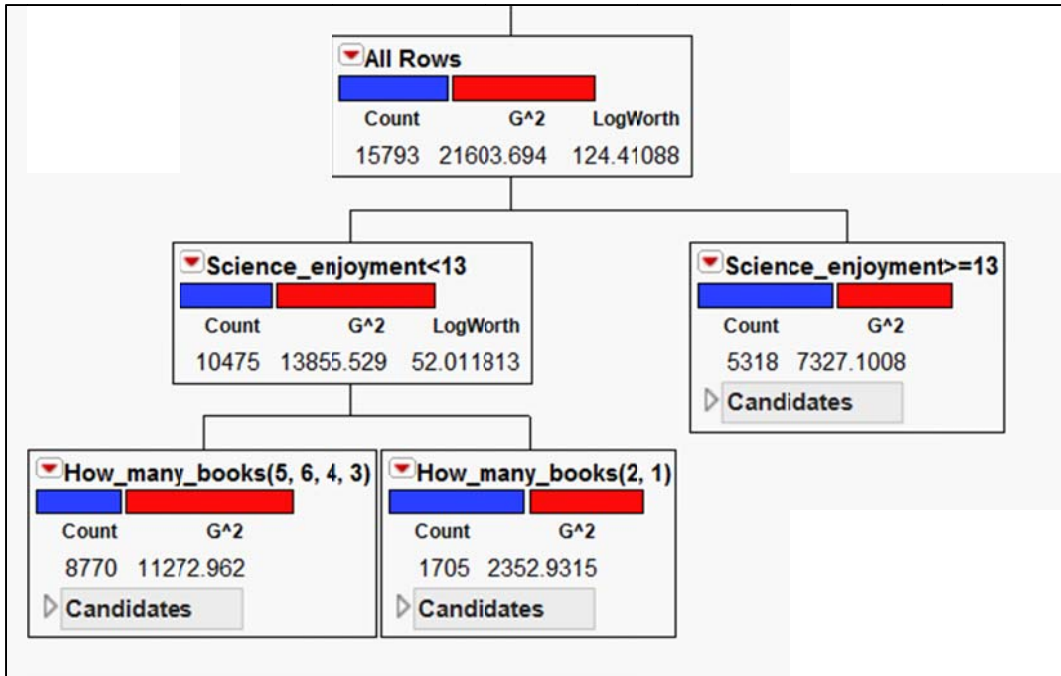
[chonghoyu@gmail.com](mailto:chonghoyu@gmail.com)

<http://www.creative-wisdom.com/computer/sas/sas.html>

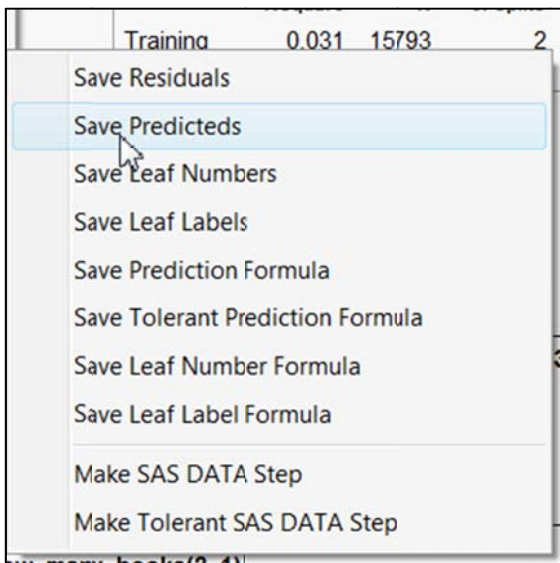
To avoid over-fitting, data mining tools in JMP have built-in resampling options such as cross-validation (CV). One of the shortcomings of CV is that the result depends on how the data set is partitioned. Specifically, the algorithm splits the data at random, and thus in each trial the researcher may see a different result. The purpose of CV is to examine the degree of model stability across various trials, and thus in principle it is good to allow fluctuations as a form of internal replication. However, sometimes the researcher may want to recreate the same output. For example, the journal editor may request modifying the same graphical output.

This write-up will illustrate how the analyst can keep a constant data partition while running a decision tree with CV. In this example the *Program for International Student Assessment (PISA)* dataset is used for demonstration. The researcher would like to identify the variables to predict proficiency (high test performance). In the first run, the researcher enters .3 in **Validation Portion** in order to hold back 70% of the sample. By doing so about 70% of the subjects is assigned into the training set and 30% is put aside for the validation set. It is important to point out that the percentage is not exact. It could be 29%, 31%...etc., because allowing fluctuation is the key for examining model stability.



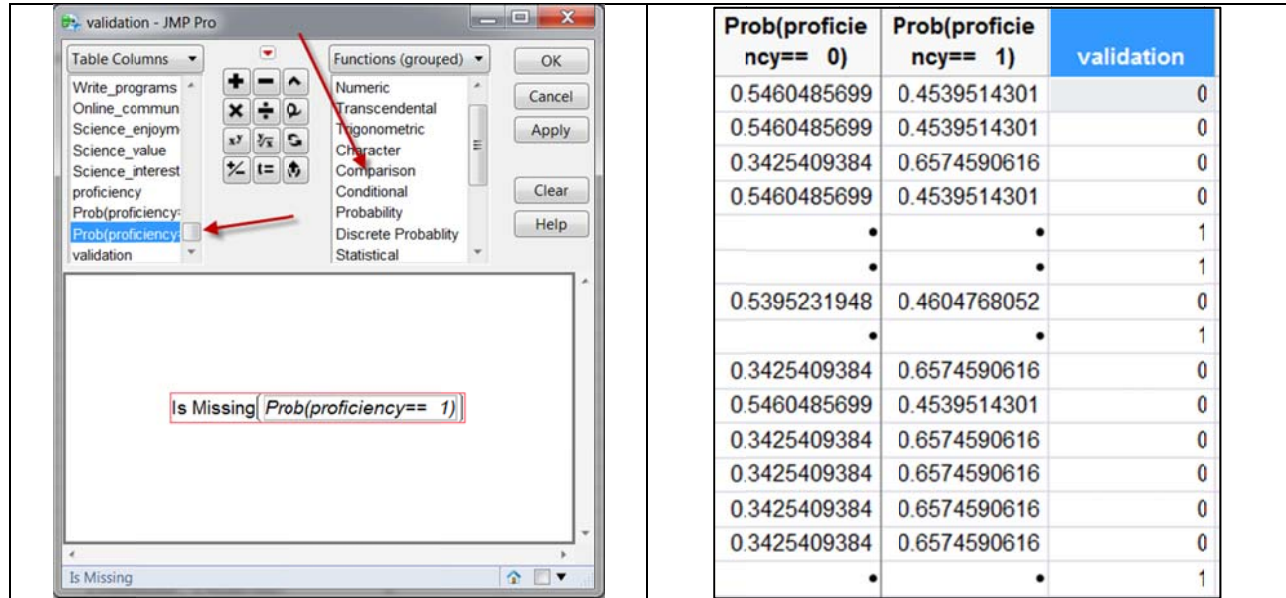


After the decision tree is created as shown above, from the red triangle choose **Save Columns**, and then choose **Save Predictions**. JMP creates two new columns in the table and each column indicates the probability of being classified as “proficient” or “not proficient”. The rows that have a numeric value are the observations used in the training set (70%) while the observations with missing values are reserved for the validation set.



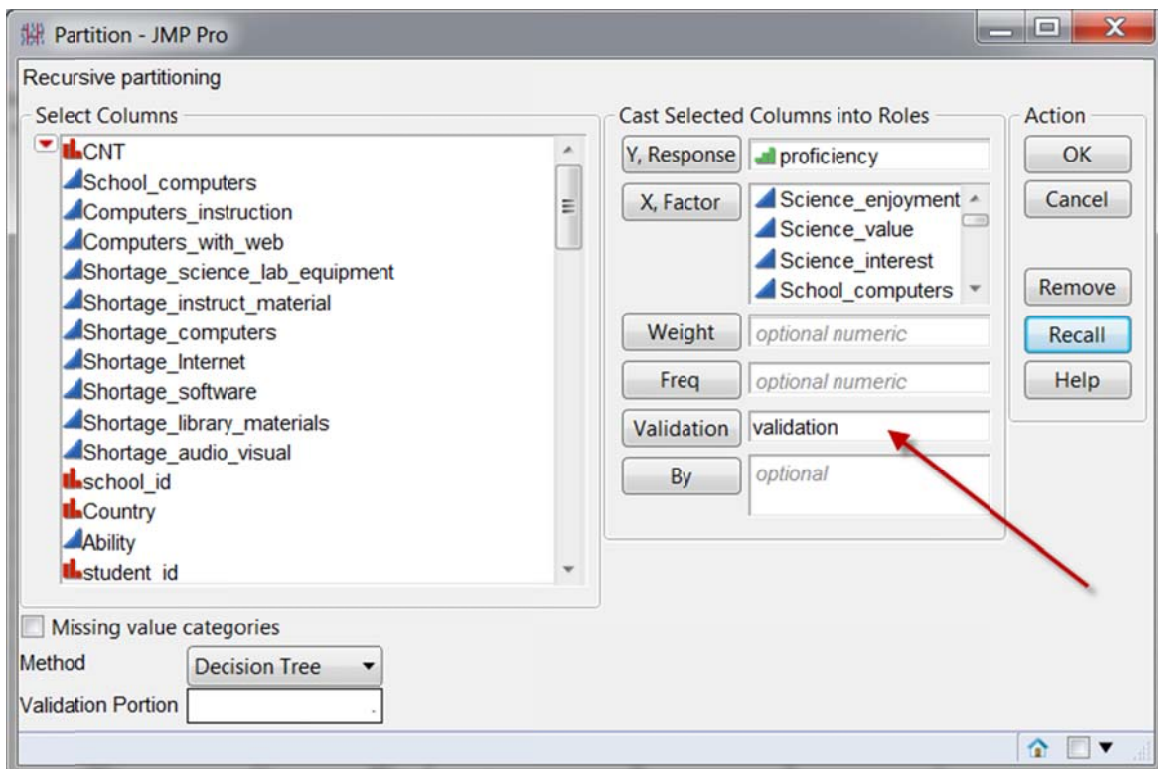
	Prob(proficie ncy== 0)	Prob(proficie ncy== 1)
	0.5460485699	0.4539514301
	0.5460485699	0.4539514301
	0.3425409384	0.6574590616
	0.5460485699	0.4539514301
	•	•
	•	•
	0.5395231948	0.4604768052
	•	•
	0.3425409384	0.6574590616
	0.5460485699	0.4539514301
	0.3425409384	0.6574590616
	0.3425409384	0.6574590616
	0.3425409384	0.6574590616
	0.3425409384	0.6574590616
	•	•

Next, create a new column and name it "Validation". Use a right click on the field name to open the **formula** window. In the formula window, first click on the variable "Prob(proficiency==1)" in **Table Columns**. Second, select **Is Missing** from **Comparison** in **Functions** and then click **OK**. Now "validation" is populated with "1" or "0" to indicate their group membership.



Prob(proficiency== 0)	Prob(proficiency== 1)	validation
0.5460485699	0.4539514301	0
0.5460485699	0.4539514301	0
0.3425409384	0.6574590616	0
0.5460485699	0.4539514301	0
.	.	1
.	.	1
0.5395231948	0.4604768052	0
.	.	1
0.3425409384	0.6574590616	0
0.5460485699	0.4539514301	0
0.3425409384	0.6574590616	0
0.3425409384	0.6574590616	0
0.3425409384	0.6574590616	0
0.3425409384	0.6574590616	0
0.3425409384	0.6574590616	0
.	.	1

When the researcher would like to re-run the same analysis using the same data partition, he can simply put "validation" into the **Validation** role.



Partition - JMP Pro

Recursive partitioning

Select Columns

- ▼ CNT
  - School\_computers
  - Computers\_instruction
  - Computers\_with\_web
  - Shortage\_science\_lab\_equipment
  - Shortage\_instruct\_material
  - Shortage\_computers
  - Shortage\_Internet
  - Shortage\_software
  - Shortage\_library\_materials
  - Shortage\_audio\_visual
  - school\_id
  - Country
  - Ability
  - student\_id

Cast Selected Columns into Roles

Y, Response: proficiency

X, Factor: Science\_enjoyment, Science\_value, Science\_interest, School\_computers

Weight: optional numeric

Freq: optional numeric

Validation: validation

By: optional

Method: Decision Tree

Validation Portion:

Action: OK, Cancel, Remove, Recall, Help