USING YOUR STORED DATA EFFECTIVELY

# DATA MINING (CUSTOMER DATA) WITH JMP

Sam Edgemon, SAS
Tony Cooper, SAS



## Audience, Purpose, Motivation

- The typical JMP user may not know about data mining*
  - even very sophisticated statisticians
- Useful to help predict outcomes (response) and behaviors
- Not often taught in school – may be changing

- New v13 (data mining) capabilities ROCK!
  - JMP can get you started building models for scoring
- And, JMP is still a very powerful visualization tool
  - Helps non-statisticians visualize the information in databases

# It Starts with Data.

### Using Dynamic Data

- Data Acquisition
  - Designed Experiments
  - Designed to give insight into a theory
  - Innovation almost always requires new data

### Using Stored Data

- Discussions on Data Mining include
  - A **process** for analyzing data using statistical models and segmentation
  - A large amount of data in a database
  - Discovering interesting patterns or relationships



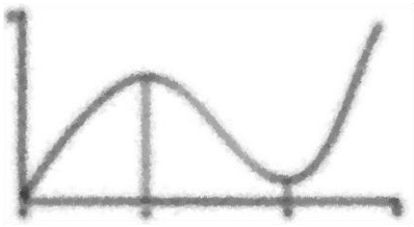


# Three Questions

1. What do you want you accomplish?
2. By what method?
3. How will you know?

*Dr. Donald J Wheeler (Quality Digest 2011)*

# Y=f(X)



Y represents the target – it's what I'm hoping to better understand by building a model.

- It's the response, typically it's the go or no go, the yes or no, they will they or won't they.
- It's the variable that's dependent on "something" else.

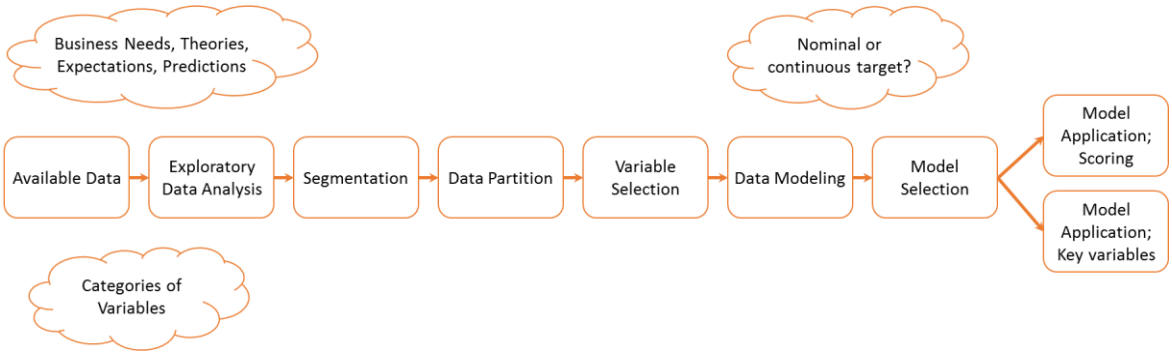
That "something" is one or more X's.

These independent variables may represent behaviors or demographics that we hope will help predict an outcome.





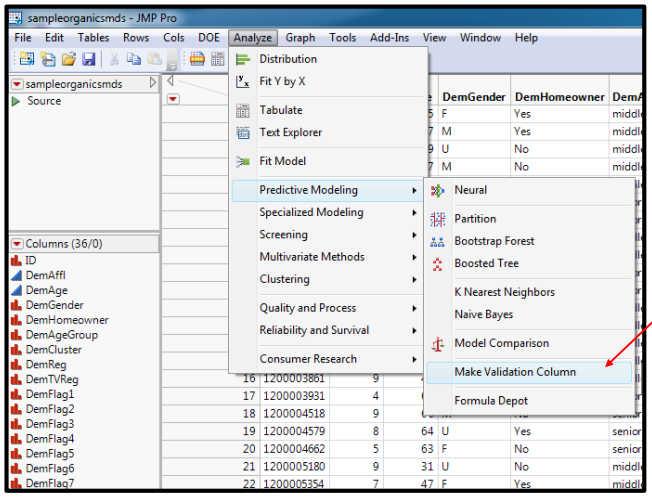
# The Process



# SCREENSHOTS FROM THE LIVE DEMO

DATA MINING WITH JMP PRO V13





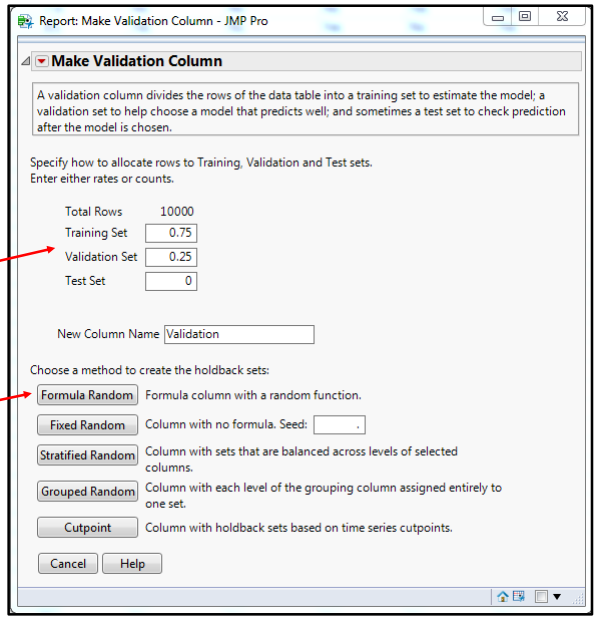

JMP makes it easy to create the “hold out” datasets necessary to validate modeling efforts.

Make Validation Column is in the Analyze > Predictive Modeling menu

You have options regarding the parameters for your training and validation datasets.

The default is set to 0.75 for the training dataset, and 0.25 for the validation dataset, but these may be changed at your discretion.

My preferred method for creating the validation column is to choose "Formula Random."

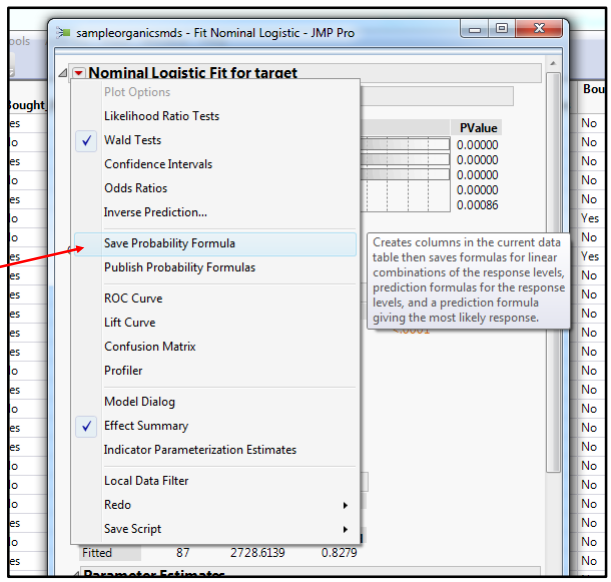


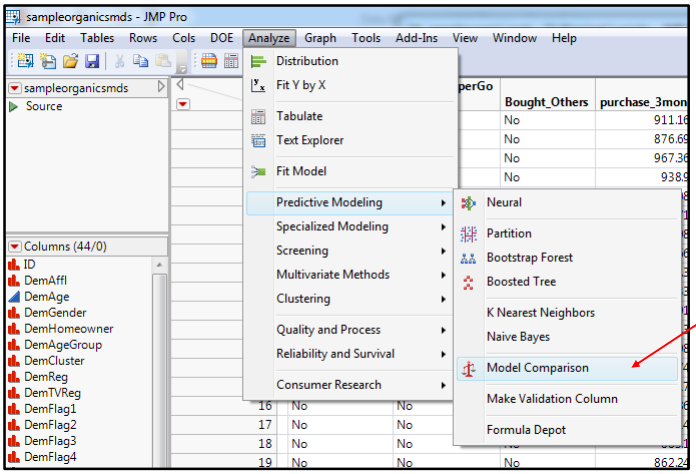



My goal is to utilize multiple modeling techniques, and I will eventually want to know which produces the best result.

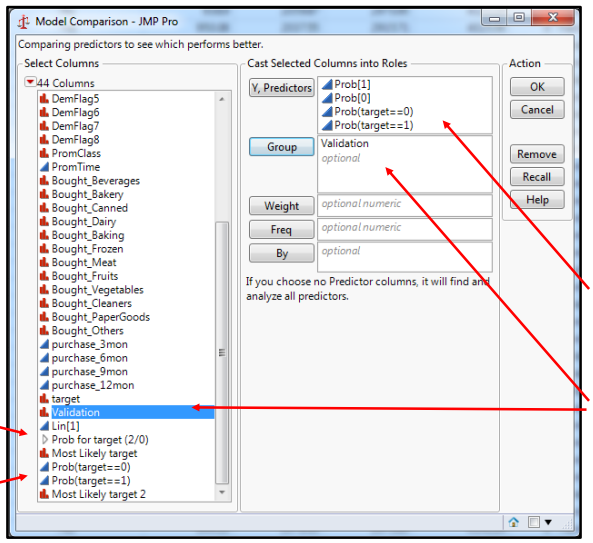
In order to take advantage of JMP's Model Comparison capabilities, I will choose "Save Probability Formula" after each model is finalized. This will create new columns in your dataset that will be utilized later to choose your best model!

Find Save Probability Formula located under the red triangle.






Model Comparison is in the Analyze > Predictive Modeling menu




The key elements regarding your model comparison will be found in your dataset.

The probability formulas for my models will have been stored.

The probability formulas for my Logistic Regression model is stored here

The probability formulas for my Bootstrap Forest are stored here

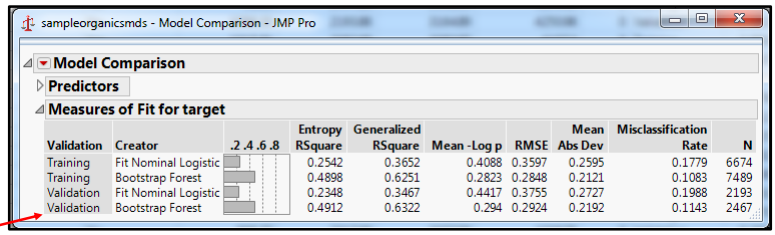
The probability formulas produced by your models will be your Y, Predictors

The Validation column will be your Group variable

Results!

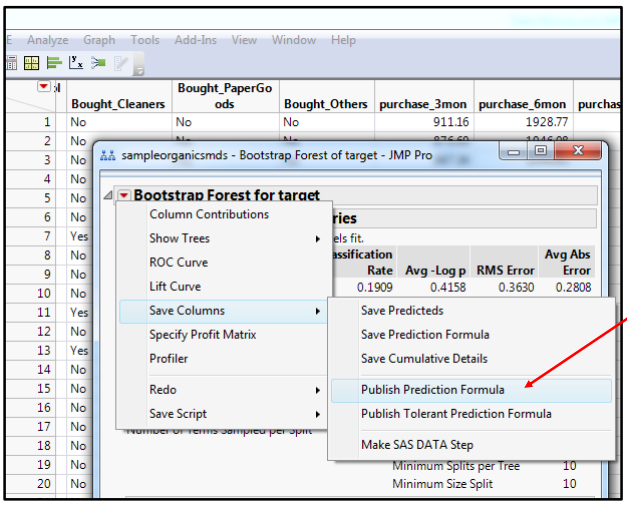
I will utilize JMP's Model Comparison output to choose the model that I will ultimately use in the business.

In this case, the Bootstrap Forest produced much better results.




| Validation | Creator | Entropy RSquare | Generalized RSquare | Mean -Log p | RMSE | Mean Abs Dev | Misclassification Rate | N |
|---|---|---|---|---|---|---|---|---|
| Training | Fit Nominal Logistic | 0.2542 | 0.3652 | 0.4088 | 0.3597 | 0.2595 | 0.1779 | 6674 |
| Training | Bootstrap Forest | 0.4898 | 0.6251 | 0.2823 | 0.2848 | 0.2121 | 0.1083 | 7489 |
| Validation | Fit Nominal Logistic | 0.2348 | 0.3467 | 0.4417 | 0.3755 | 0.2727 | 0.1988 | 2193 |
| Validation | Bootstrap Forest | 0.4912 | 0.6322 | 0.294 | 0.2924 | 0.2192 | 0.1143 | 2467 |





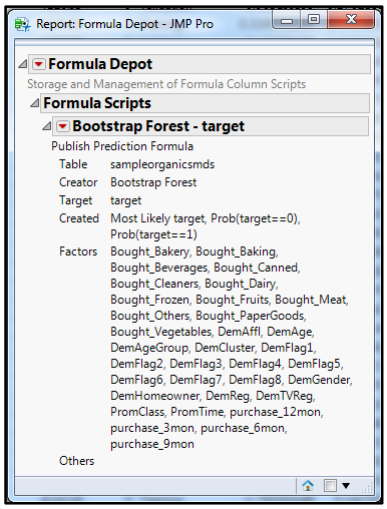

Action!

After using the Model Comparison tool to select the model, I go back to the model and from the red triangle, I go Save Columns > Publish Prediction Formula.

The Formula Depot will be called with your information set up to be run or saved.

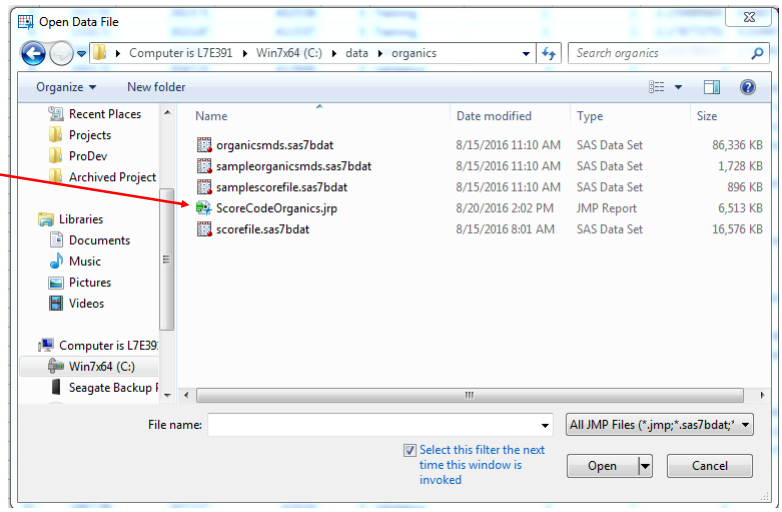
The Formula Depot sits ready with the models parameters stored for future use!

I will close and save this Formula Depot to score future datasets.




I saved my “score code” as ScoreCodeOrganics.jrp with the intention of receiving a dataset each week.

ScoreCodeOrganics will create a new column representing the likelihood of a customer responding in a particular way.



# THANKS!

Sam Edgemon
sam.edgemon@sas.com

Tony Cooper
tony.cooper@sas.com