SAS® Modeling Best Practices

Melodie Rush
Global Customer Success Principal Data Scientist

Connect with me:
LinkedIn: https://www.linkedin.com/in/melodierush
Twitter: @Melodie_Rush

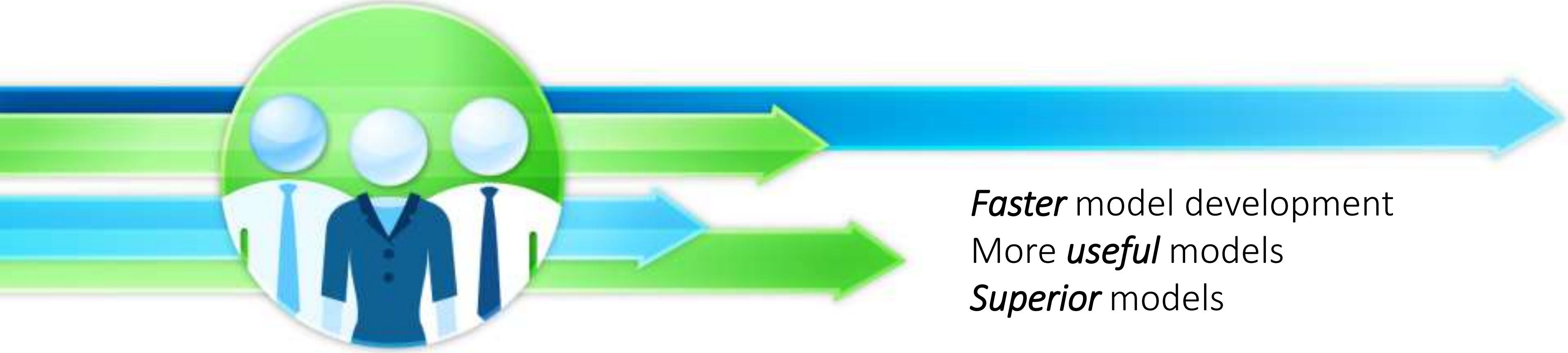
Agenda

SAS Enterprise Miner

Best practices for creating a predictive model

- Background and General Guidance
- Data Construction
- Model Development and Delivery

Best practices to help you meet and exceed your goals



Faster model development
More ***useful*** models
Superior models

Best Practices

Disclaimers

- The choice of “Best Practices” is highly subjective.
- Certain suggested practices may not be suitable for a particular situation.
- It is the responsibility of a data mining practitioner to critically evaluate methods and select the best method for a particular situation.
- This presentation represents the opinions of those who contributed.

Background

Analytics Cycle and the modeling Process

Why use Predictive Modeling?

To Turn increasing amounts of raw data into useful information

Descriptive

Clustering (Segmentation)

grouping together similar people, things, events

- Transactions that are likely to be fraudulent, Customers that are likely to have similar behaviors.

Associations

affinity, or how frequently things occur together, and sometimes in what order

- Customers who purchase product A also purchase product B

Predictive Models

Classification models

predict class membership

- 0 or 1: 1 if person responded; 0 otherwise
- Low, Medium, High: a customer's likeliness to respond

Regression models

predict a number

- $217.56 – Total profit, expense, cost for a customer
- 37 – The number of months before a customer churns

The Goal? Scoring!

- Scoring is the act of applying what we’ve learned from data mining to **new cases**.
- Keep this goal in mind and use it to help formulate the questions and the data needed for data mining and scoring.

Example

Developing a Classification Model

- Models are developed using historical data in which the **behavior is observed or known**.

Indicates the behavior was observed in this subject

- Information about each subject, in this case an individual, is used as inputs to the model to see how well the model can distinguish between the people who exhibit the behavior and those who do not. For example, age, gender, previous behaviors, etc.

Why?

- Consider a group of subjects whose relevant behavior is unknown.
- The same information is available for each of these subjects (age, gender, etc.) as is available for the individuals with known behavior.
- We would like to know **which individuals are most likely to have the relevant behavior**.

How?

- The output of a predictive classification model is typically an equation. Models are applied to new cases to calculate the **predicted behavior** through a process called **scoring**.
- **Scoring**, using the equation, calculates each subject's *likelihood to have the relevant behavior*. (It also calculates the likelihood to *not* have the behavior.)

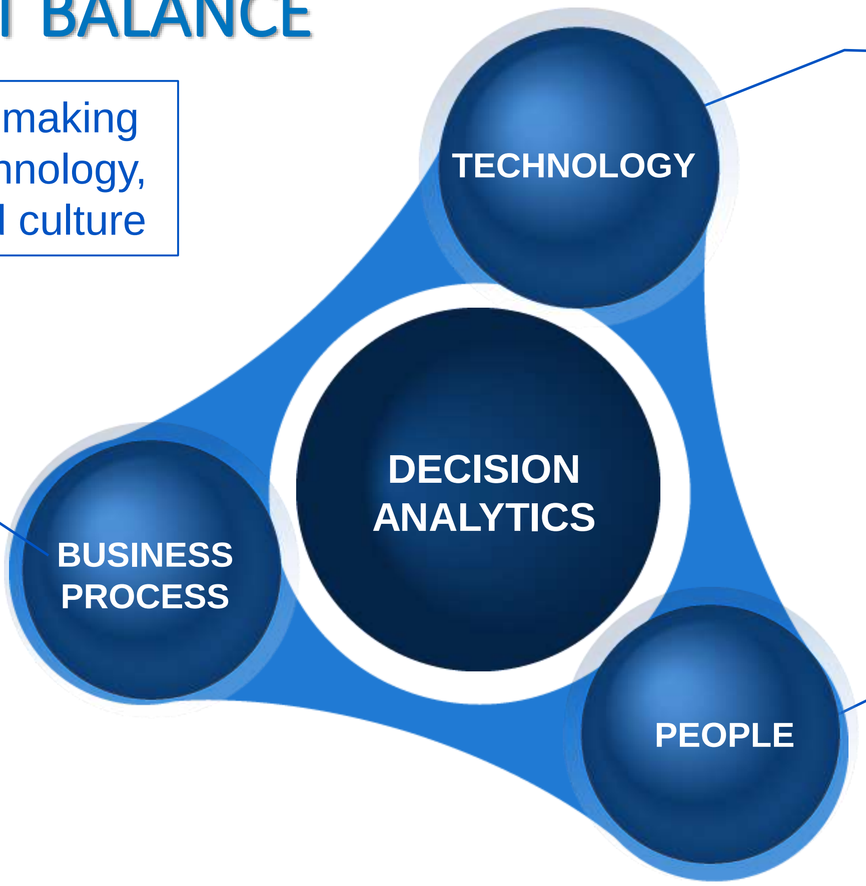
General Guidance

Analytics Cycle and the Modeling Process

ITS ALL ABOUT BALANCE

Fact-based decision making requires the right technology, talent, processes and culture

- Continuous Process Improvement
- Planning
- Project methodology
- Standards



- BI reporting
- Web portals / dashboards
- Information management
- Problem-specific business solutions
- Predictive analytics
- Hardware

- Vision & Leadership
- Team composition
- Enterprise authority

Lifecycle Best Practice

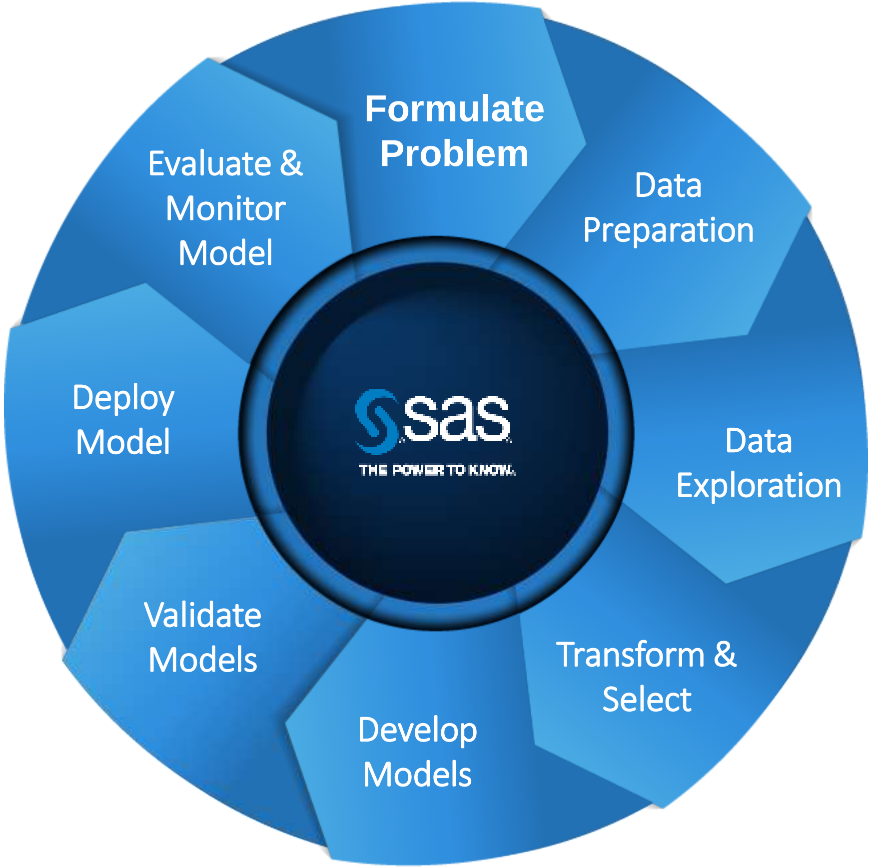
Involve all the relevant people/roles

BUSINESS MANAGER

Domain Expert
Makes Decisions
Evaluates Processes & ROI

BUSINESS ANALYST

Data Exploration
Data Visualization
Report Creation



DATA MINER DATA SCIENTIST

Exploratory Analysis
Descriptive Segmentation
Predictive Modeling
Model Validation & Registration

IT/SYSTEMS MANAGEMENT

Model Validation
Model Deployment
Model Monitoring
Data Preparation

sas

Best Practice

Use the Technology and Method the Fits the Job

Every tool and method has advantages and disadvantages.

Whenever possible, select the tool or method that balances *long-term* goals for the *entire* process.

Best Practice

Begin with the End in Mind

Best Practice

Begin with the End in Mind

- *What* is the overarching strategic objective/initiative?
- *How* will the model be used?
- *How* will it be put into production?
- *Who* will be affected by the use of the model?
- *Who* needs to be convinced of the value of the model?
- *When* will the model be used?

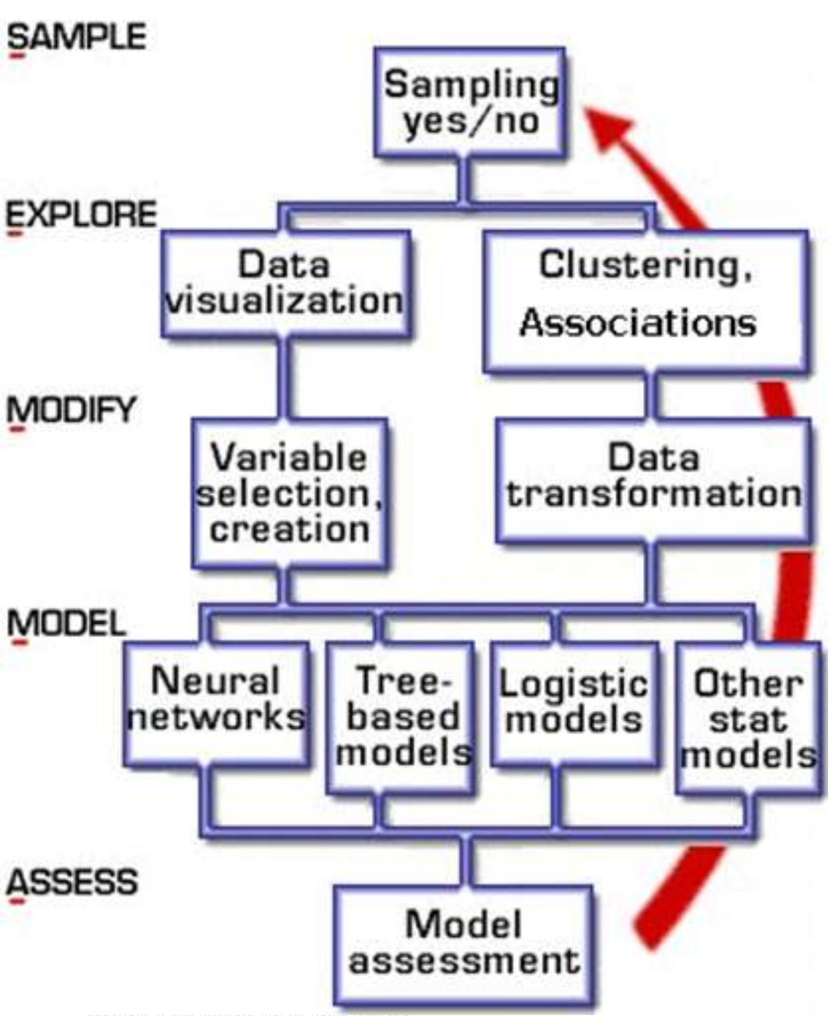
Best Practices

Business considerations Before you model

- Thoroughly understand the business/marketing objectives
- Detail the precise (planned) usage for the output
- Define the target variable (the outcome being modeled / predicted)
- Formulate a theoretical model: $Y = f(X_1, X_2, \ldots)$ ← fill-in the likely X's

BEST PRACTICE

SEMMA Process for Model Development



Best Practice

Modeling Approach

1. **Sample** → training set(s), validation set(s), holdout test set
2. **Explore** → min, max, mean, median, missing values, levels (categorical cardinality)
3. **Modify** → filtering outliers, reducing cardinality, correcting multicolinearity, imputations, non-linear transformations

Best Practice

Modeling Approach

4. **Model** → variable selection, various model formulations, iterative cycle, insights & client reviews
5. **Assess** → performance criteria and review

Best Practice

Modeling approach (Continued)

6. Final Assessment & Testing
7. Profile characteristics & indicators
8. Document results
9. Prepare (production-ready) data collection and score code
10. Monitor model performance

Developing the Data

Best Practices

Optimizing Data

Determining Data

Selecting Target

Preparing Variables

Determining Data

Best Practices

Technical Considerations Before Modeling

- Brainstorm all potential input data elements
- Identify source systems, specific data fields, availability/priority/level-of-effort of data
- Finalize data to be collected

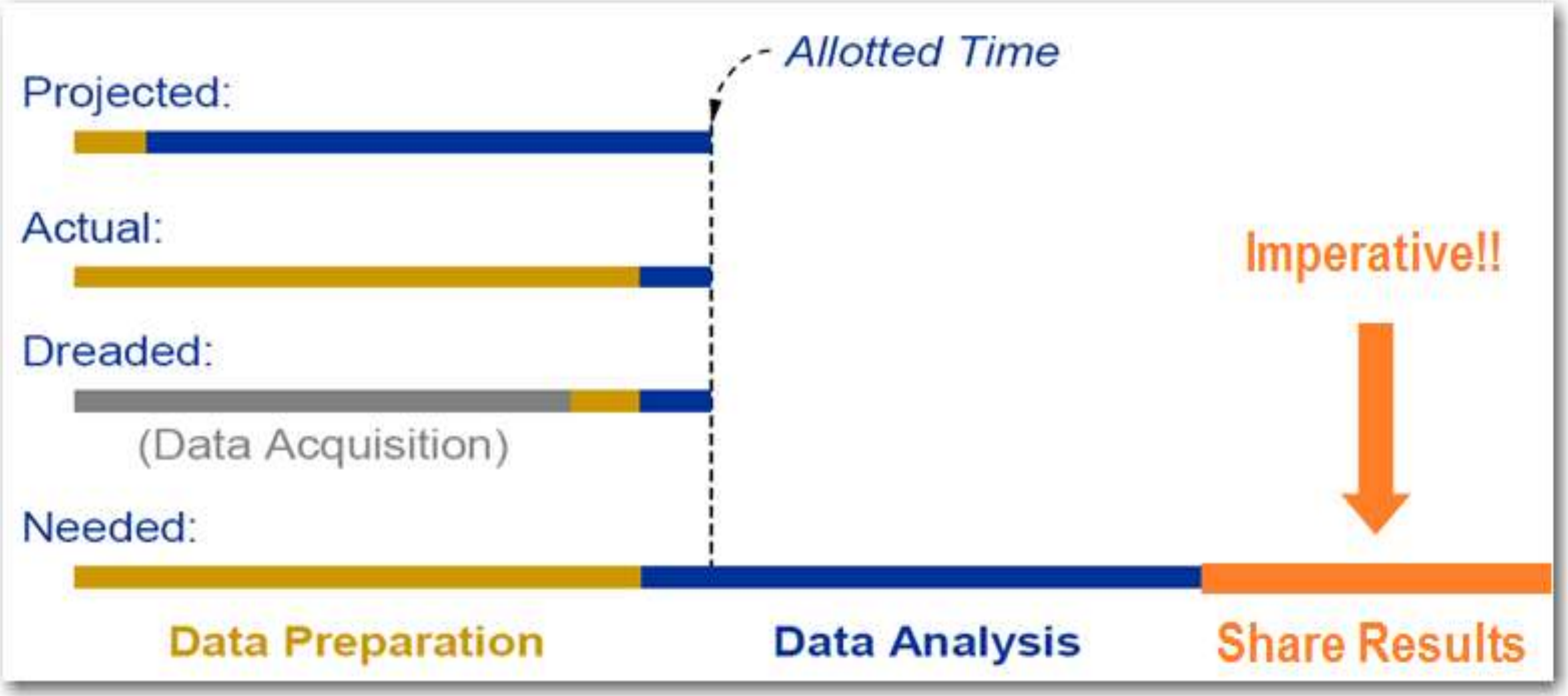
Best Practices

Technical Considerations Before Modeling

- Formulate structure and layout of modeling dataset to be built
- Devil-in-the-details: filters, timeframe of history, etc…
- Build modeling dataset

Best Practice

Allow sufficient time for all aspects



Best Practice

Sample

- (Over) Sampling
- Partitioning
- Decisioning

Sample

Sample

To Sample or Not?

- Sampling is a valuable tool that can be used to great effect.
- If computing resources are no object, it's possible to use all data.
- When resource constrained, try increasing sample sizes as model development progresses.
- When model is nearly finalized, try different seeds for samples to ensure model stability.

Sample

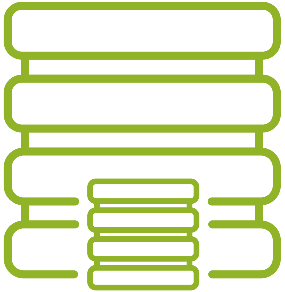
What About Oversampling?

- **It depends.**
- Frequently one needs to oversample in order to allow algorithm(s) to identify effect, especially with rare targets.
- Only oversample as much as you need to in order to obtain a model that makes sense from a business perspective. This is **highly subjective**.

Adjusting for Oversampling

Why?

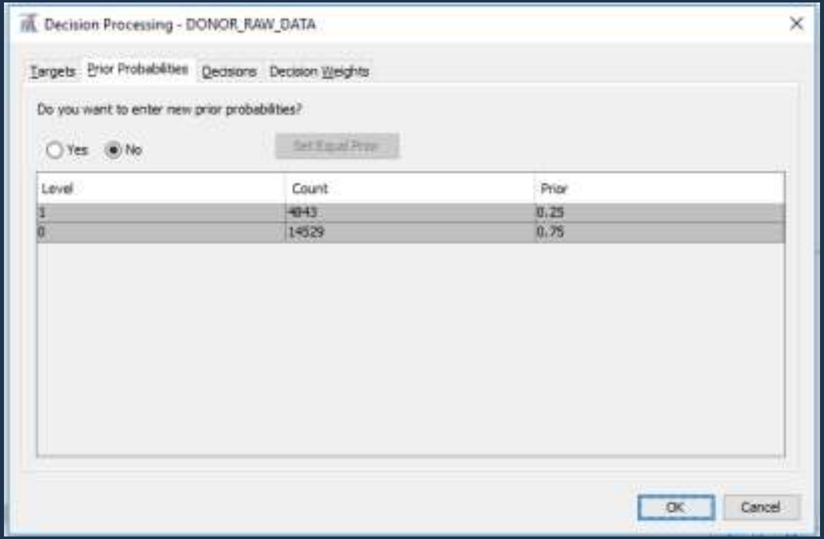
- Prediction estimates reflect target proportions in the training sample, not the population from which the sample was drawn.
- Score Rankings plots are inaccurate and misleading,
- Decision-based statistics related to misclassification or accuracy misrepresent the model performance on the population.



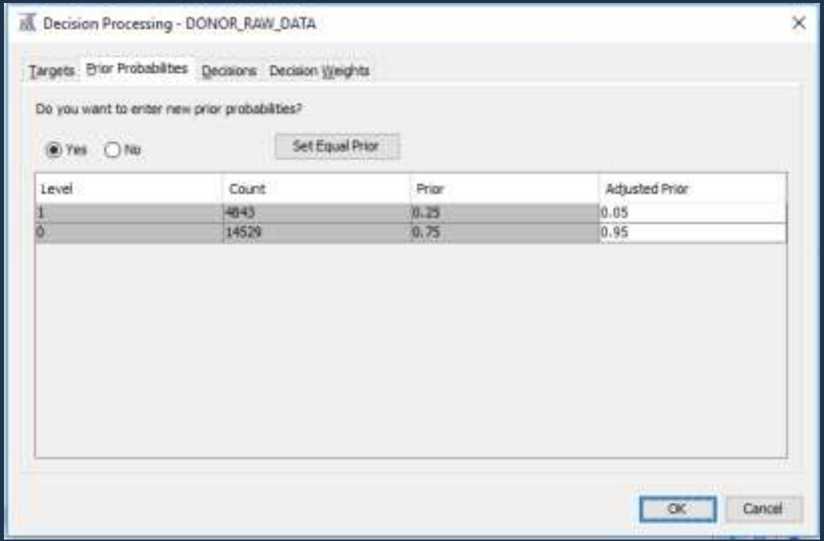
Adjusting for Oversampling

Prior Probabilities

Before



After



Adjusting for Oversampling

Model Comparison

Before

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Selection Criterion: Valid: Average Squared Error
Y	Neural	Neural	Neural Net...	TARGET_B	0.18275
	Reg	Reg	Regression	TARGET_B	0.183045
	Tree	Tree	Decision Tr...	TARGET_B	0.184104

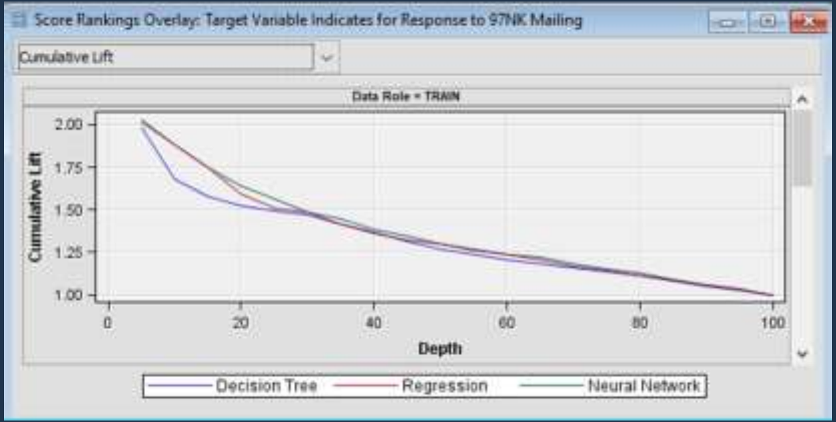
After

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Selection Criterion: Valid: Average Squared Error
Y	Neural	Neural	Neural Net...	TARGET_B	0.18275
	Reg	Reg	Regression	TARGET_B	0.183045
	Tree	Tree	Decision Tr...	TARGET_B	0.184104

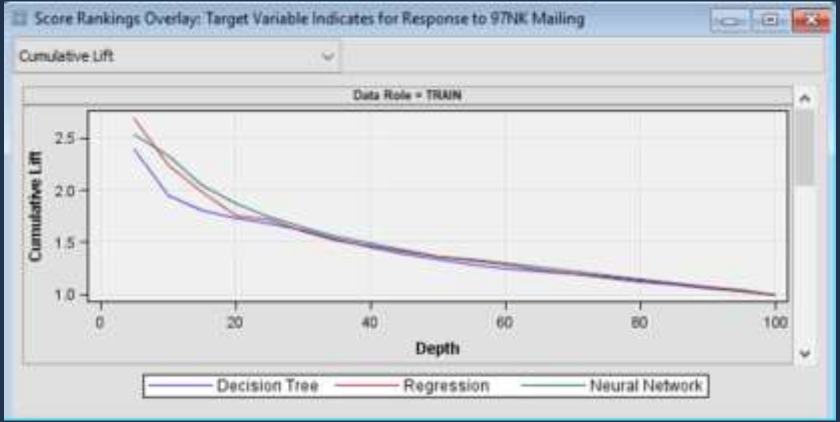
Adjusting for Oversampling

Cumulative Lift

Before



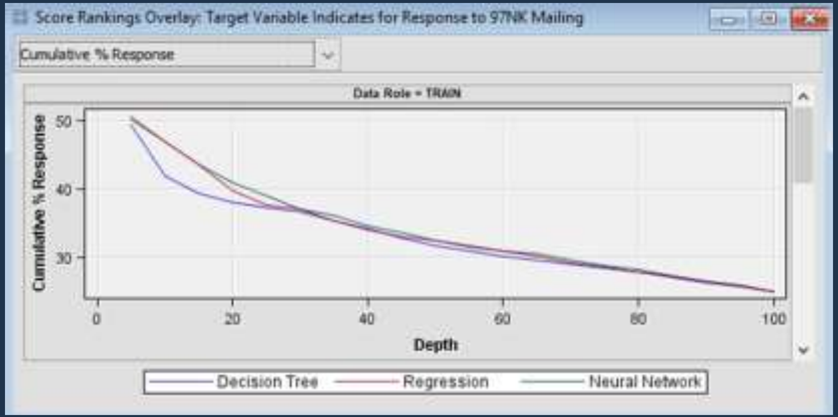
After



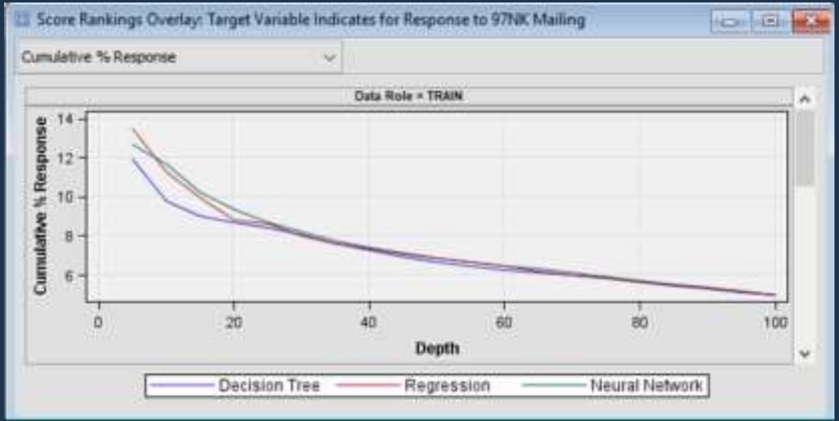
Adjusting for Oversampling

Cumulative % Response

Before



After



Decisions

Incorporating Priors

- Before fitting model
  - Decision Profile
- After fitting model
  - Decision Node

Partitioning

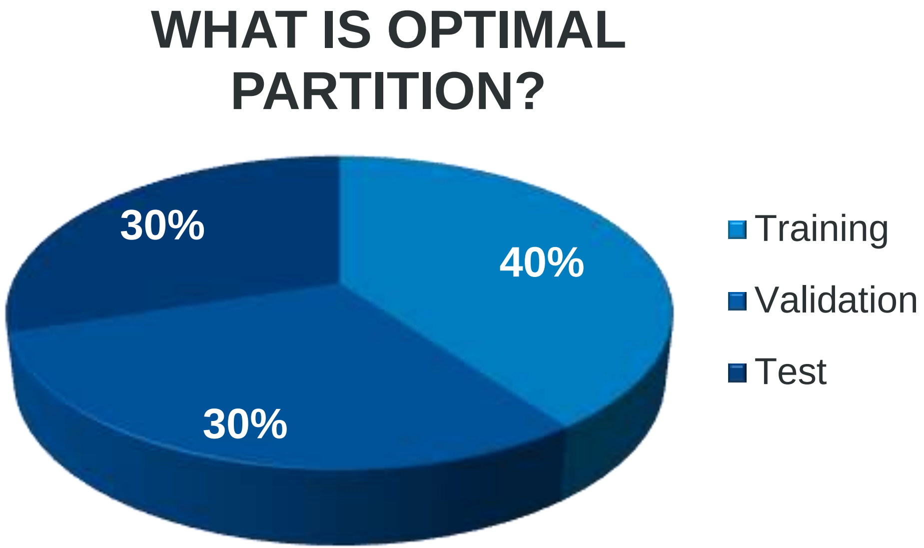
SAMPLE

Data Partitioning

PARTITION	ROLE
Training	Used to fit the model
Validation	Used to validate the model and prevent over-fitting
Test	Used to provide unbiased estimate of model performance

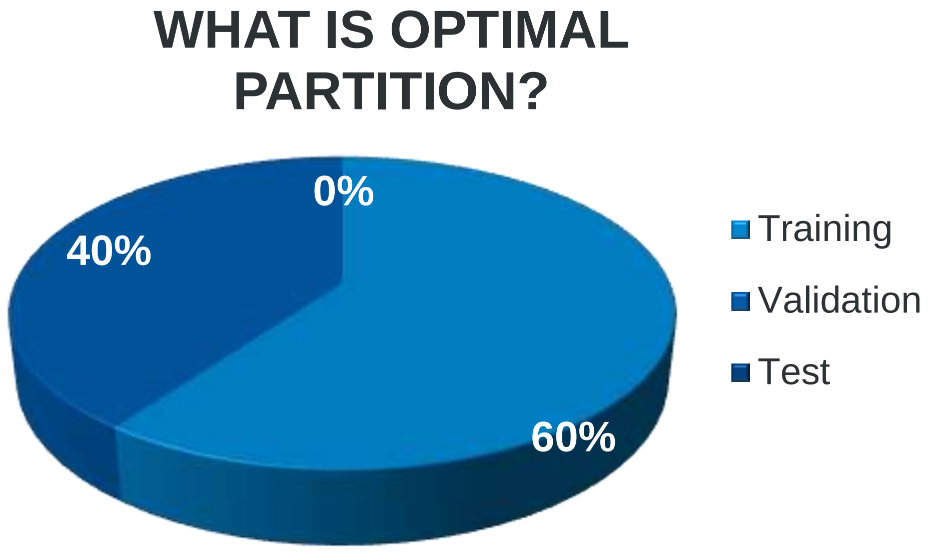
Sample

SAMPLE: Data Partitioning



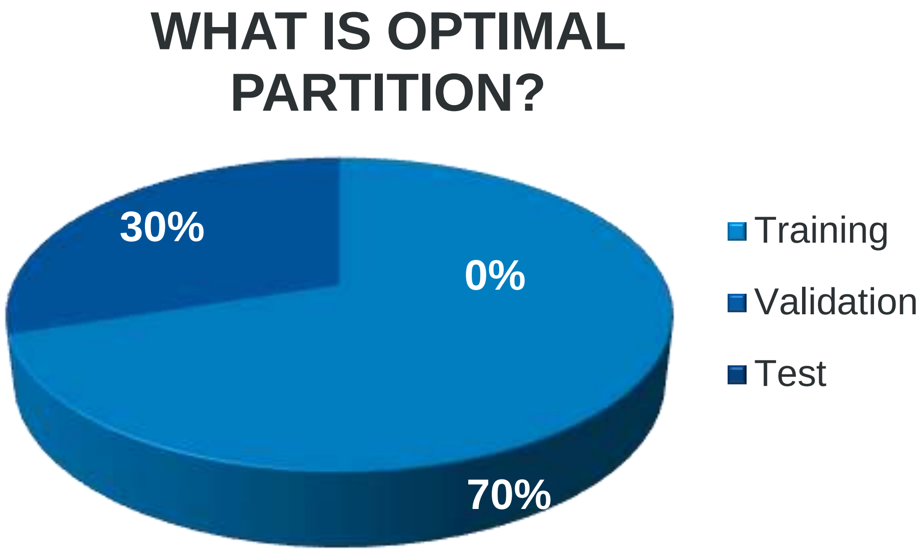
Best Practice

SAMPLE: Data Partitioning



Best Practice

Sample: Data Partitioning



Sample

Data Partitioning Considerations

- How much data is available?
- Is an unbiased measure of model performance required?
  - Should test data be in-sample or out-of-sample?
- How many test samples are needed? (e.g. different time periods, different geographies, etc.)
  - When should test data be used in the process?

Best Practice

Data Partitioning

- Percentages: frequently used percentages are 50/50/0, 60/40/0 and 70/30/0 with a completely separate Test partition.
- Do not bring Test data into process until model is complete. It should not influence modeling process, merely used to report performance.
- Multiple Test data can be used – consider how model will be deployed and create representative samples.

Decisioning

Weighting Your Decisions

- Expected Profit
- Decision Boundaries

Understanding expected profit

- Consider this game
  - Flip a fair coin one time
  - If it is heads, you win $10.00
  - Cost of playing one time is $1.00

Do you want to play this game?

Understanding expected profit

- Consider this game
  - Flip a fair coin one time
  - If it is heads, you win $10.00
  - Cost of playing one time is $1.00

$$E(Profit) = 0.5 * (10 - 1) + 0.5 * (-1)$$

$$= \quad 4.50 \quad + \quad (-0.50) \quad = \; \mathit{4.00}$$

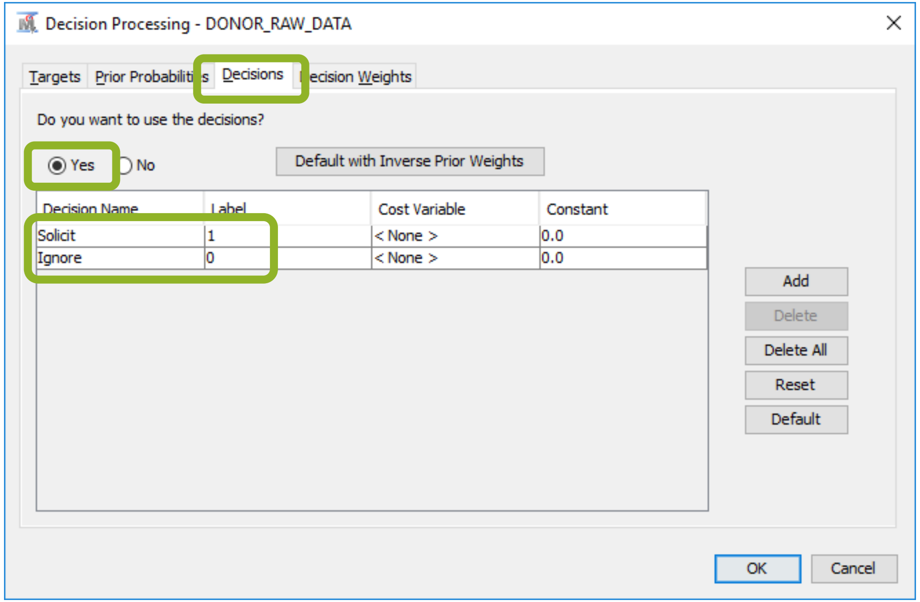
Decision Theory

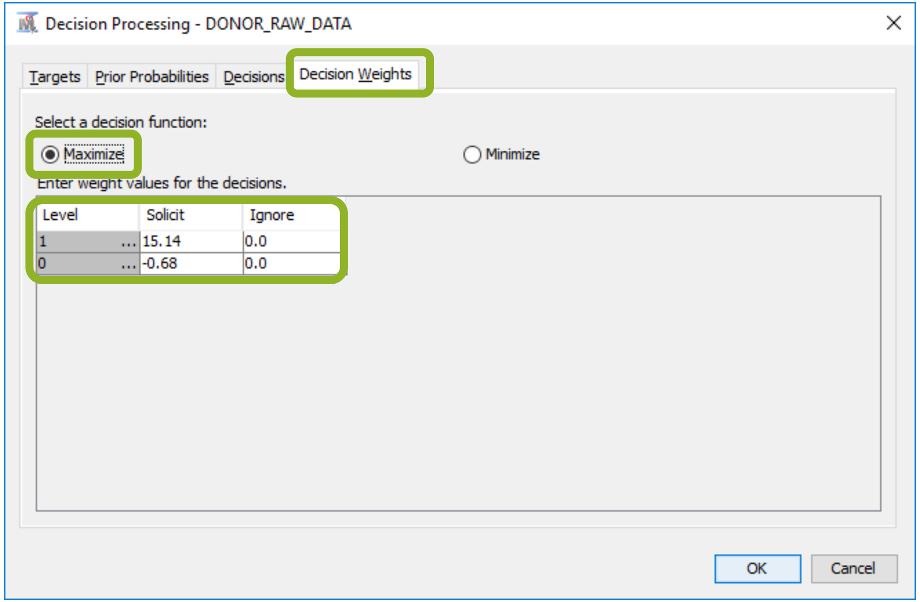
What is it?

- Decision Theory is an aid to making optimal decisions from predictive models.
- Each target outcome is matched to a particular decision or course of action.
- A profit value is assigned to both correct and incorrect outcome and decision combinations.
- The best model is selected based on maximizing profit or minimizing cost.

Decisions

Combining the Decisions with Weights





Adjusting for Oversampling

Model Comparison

Before

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Selection Criterion: Valid: Average Squared Error
Y	Neural	Neural	Neural Net...	TARGET_B	0.18275
	Reg	Reg	Regression	TARGET_B	0.183045
	Tree	Tree	Decision Tr...	TARGET_B	0.184104

After Prior Probability Adjustment

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Selection Criterion: Valid: Average Squared Error
Y	Neural	Neural	Neural Net...	TARGET_B	0.18275
	Reg	Reg	Regression	TARGET_B	0.183045
	Tree	Tree	Decision Tr...	TARGET_B	0.184104

Adjusting for Oversampling

Model Comparison

Before

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Selection Criterion: Valid: Average Squared Error
Y	Neural	Neural	Neural Net...	TARGET_	0.18275
	Reg	Reg	Regression	TARGET_	0.183045
	Tree	Tree	Decision Tr...	TARGET_	0.184104

After Applying Profit and Costs

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Selection Criterion: Valid: Average Profit for TARGET_B
Y	Reg	Reg	Regression	TARGET_B	0.164931
	Neural	Neural	Neural Net...	TARGET_B	0.161249
	Tree	Tree	Decision Tr...	TARGET_B	0.145189

Selecting Target

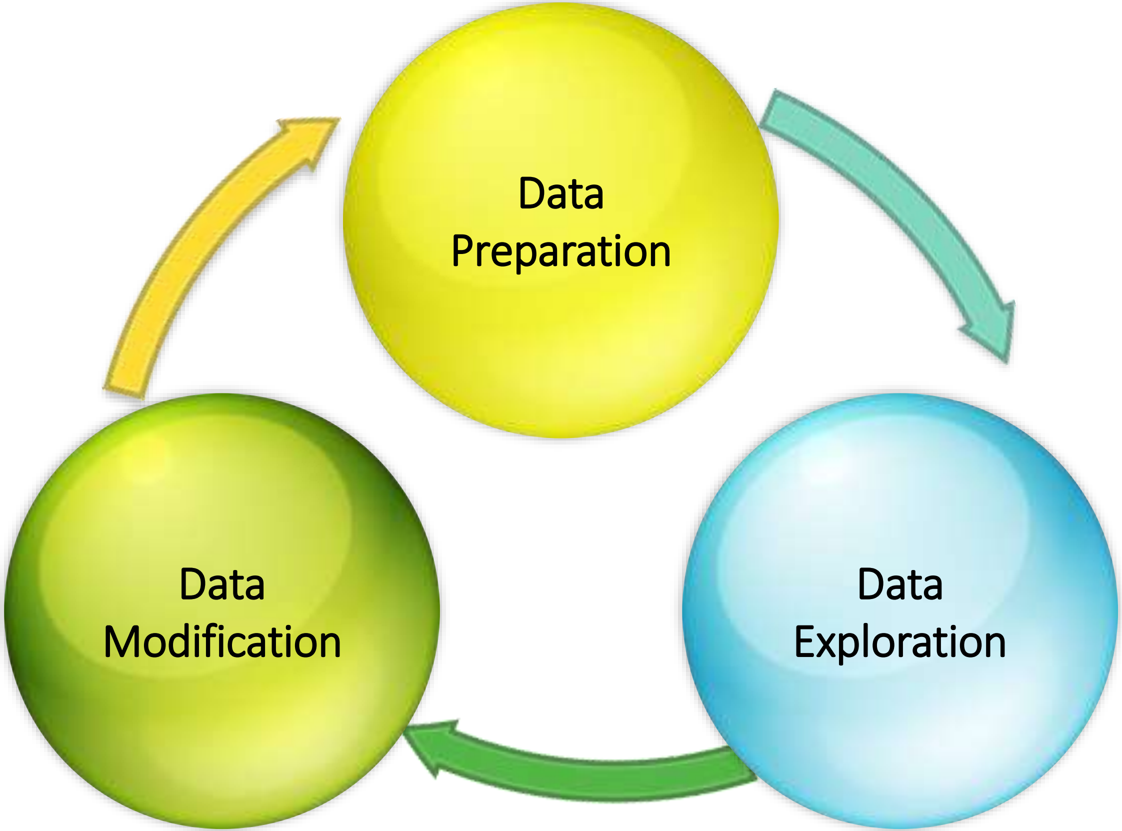
Choosing your target

- Choosing the Target
- Response vs. Propensity
- Number of Models

Preparing Data

EXPLORE & MODIFY

Iterative Relationship with Data Preparation



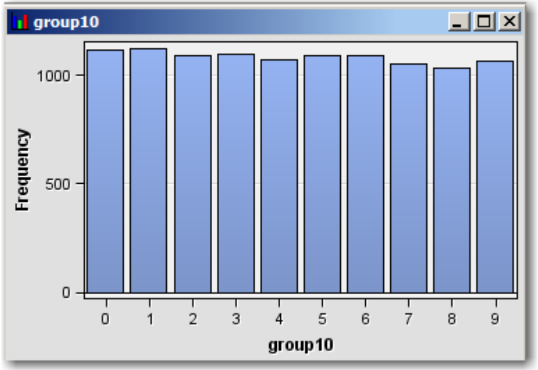
Best Practice

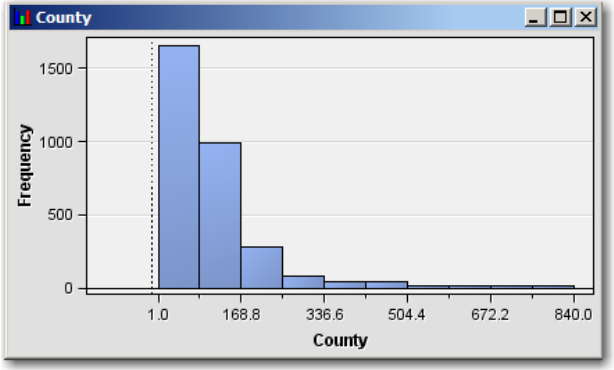
Explore & Modify: Getting the Most out of Data

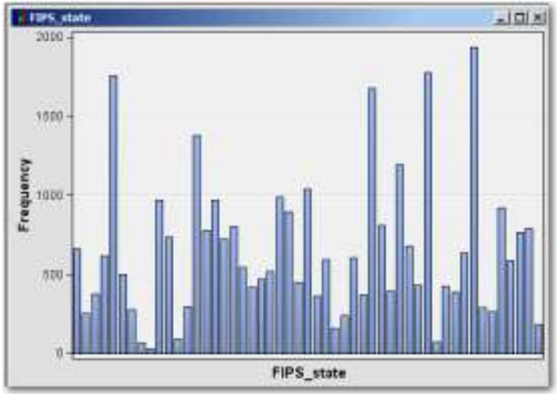
- Once you have an analytics-ready table:
  - Examine *Categorical* Variables
  - Examine *Continuous* Variables
  - Explore *Missing* Values
  - *Cluster* Variables

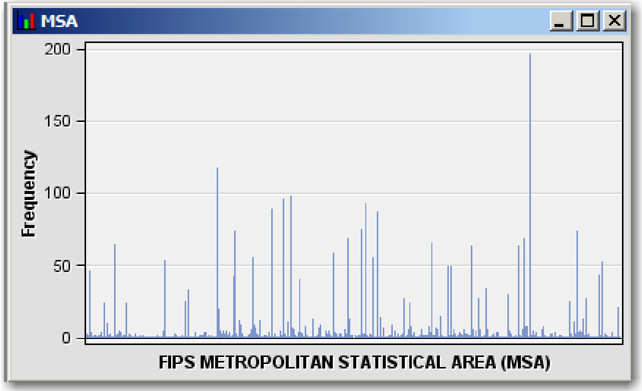
Explore & Modify

Categorical Variables



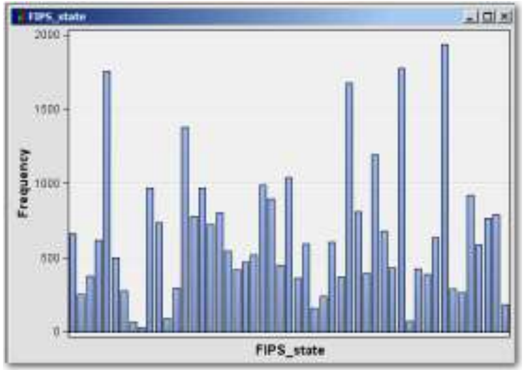


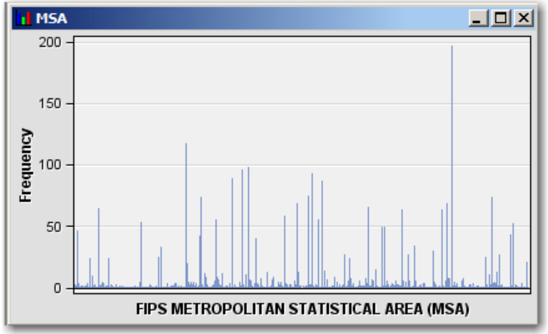




Explore & Modify

Categorical Variables



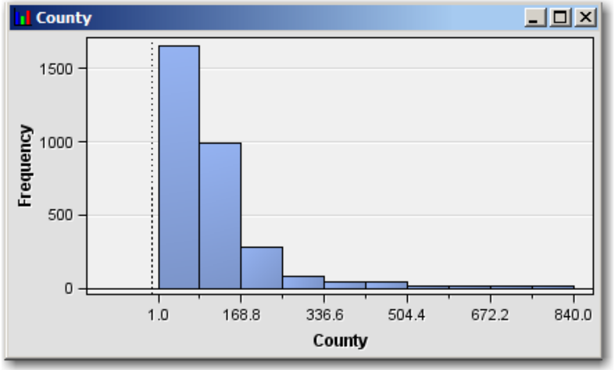


Too many overall values

- Is there a higher level of a hierarchy that could be used instead?
- Can this be represented by a group of variables with fewer values?
  - Example: **Zip Codes** alternatives
  - MSA or state
  - Geographic, demographic, economic status

Explore & Modify

Categorical Variables

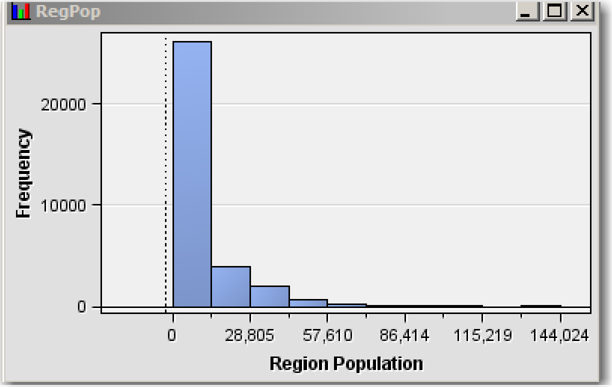


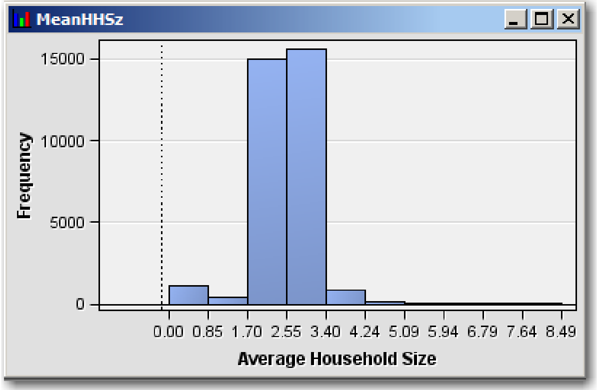
Levels that rarely occur

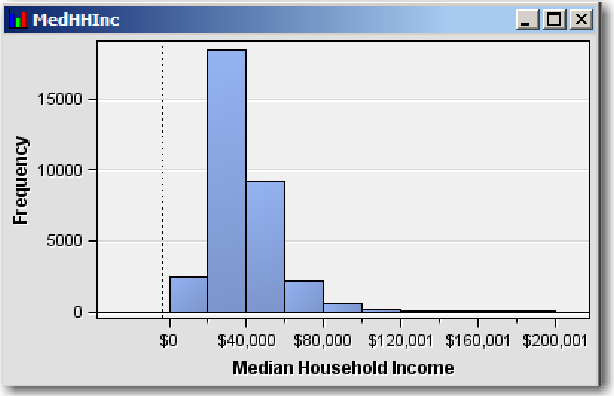
- Group infrequently occurring values together as "other"
- Judiciously combine a less frequently occurring level with a more frequent one where it makes business sense
- Consider a less granular level of a hierarchy

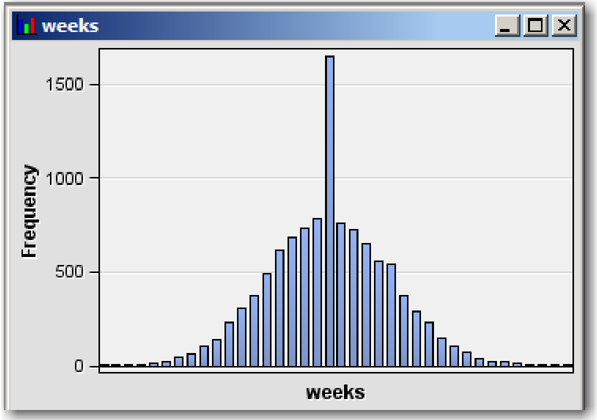
Explore & Modify

Continuous Variables



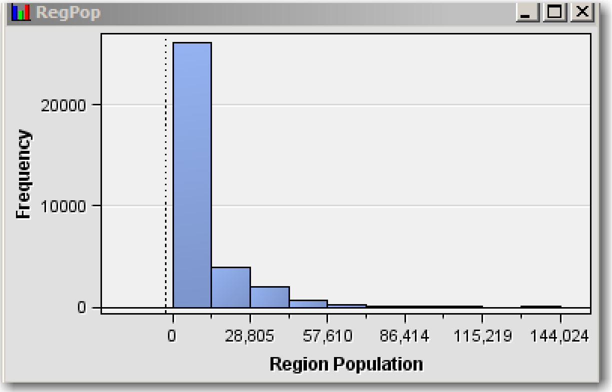






Explore & Modify

Continuous Variables



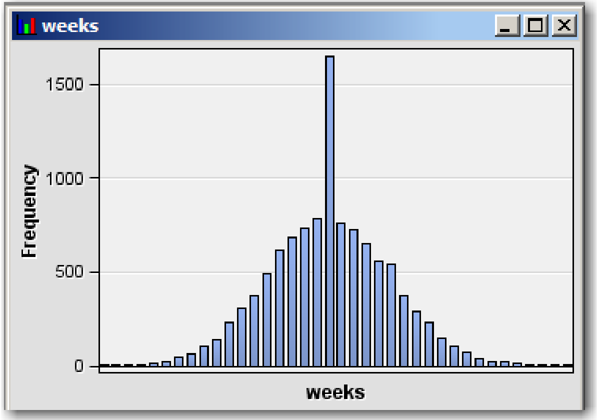
Extremely skewed predictors

- Consider transformations that stabilize variance and generate more support across the range of values
- Consider binning transformation with appropriate number of bins to enable each portion of the ranges to be weighed appropriately

Explore & Modify

Continuous Variables

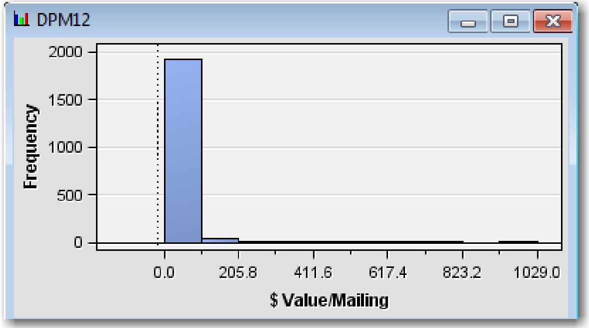
Spike and a Distribution



- Consider creating two variables from the original
  - Flag variable to indicate whether value is in the spike
  - Variable from the values of the predictors in the distribution
    - Set values at spike to missing and impute

Explore & Modify

Continuous Variables



One level that almost always occurs

- Consider a new variable that is a binned version
- Consider whether it’s sufficient to create only a binary indicator

Explore & Modify

Missing Data

- *Why* is data missing?
- Are there *patterns* to the missing data within or across variables?
- *Imputation methods* to consider
- *Indicator variables*

Explore & Modify

Variables for Clustering

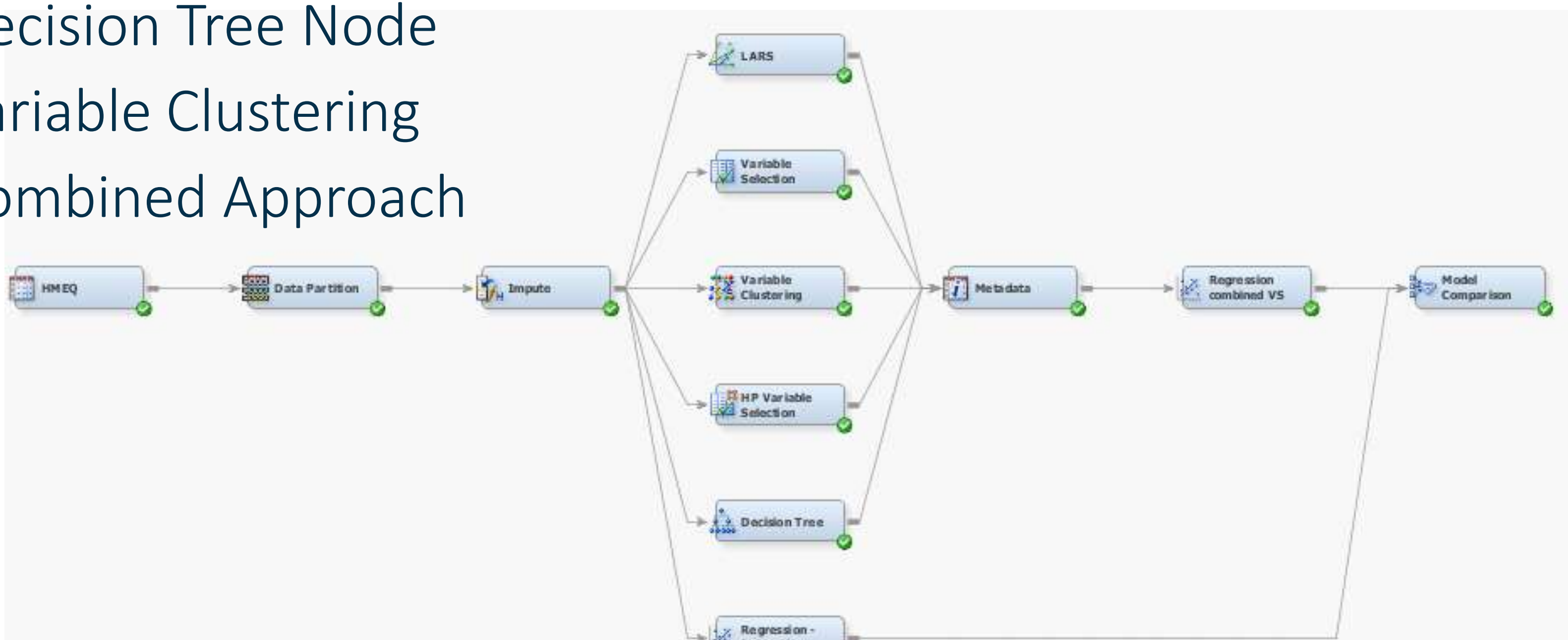
- There is no single answer for clusters
- Design clusters and profiles around themes using smaller set of related variables

Selecting Variables

Explore & Modify

Variable Selection/Reduction Techniques

- Stepwise Regression
- Variable Selection Node
- Decision Tree Node
- Variable Clustering
- Combined Approach



Best Practices

Optimizing Data

Determining Data
Selecting Target
Preparing Variables

Developing & Delivering the Model

Model & Assess

Delivering the Model

- *Developing* Your Model
- *Choosing* a Model
- *Deploying* the Model

Developing the Model

MODEL

Model Development

- Regression
- Decision Trees
- Neural Networks
- Ensemble
- Random Forest
- Something Else?

BEST PRACTICE

Model Development

- Try various techniques and combinations of techniques.

Choosing a Model

Best Practices

Model Selection

- Evaluate model metrics
- Consider business knowledge
- Recognize constraints

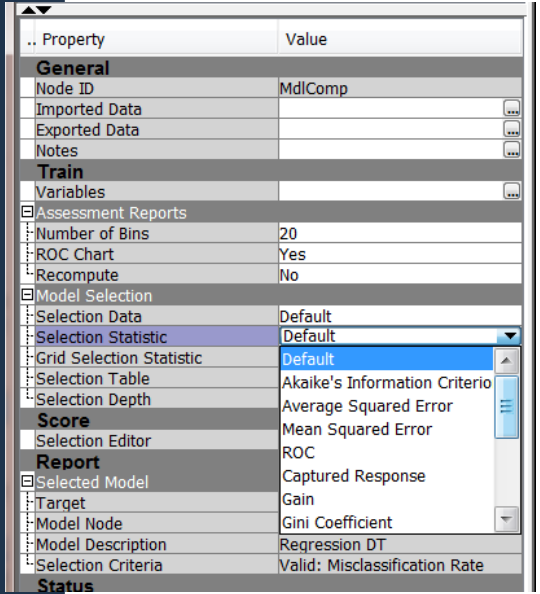
How?
Model Selection Criteria

- Decisions/Assessment
  - Accuracy/Misclassification
  - *Profit/Loss*
  - *Inverse prior threshold*
- Estimates
  - Average squared error
  - SC (SBC or BIC)
- Rankings
  - ROC Index
  - Gini coefficients

Validation Fit Statistic Direction

Prediction Type	Validation Fit Statistic	Direction
Decisions	Misclassification	smallest
	Average Profit/Loss	largest/smallest
	Kolmogorov-Smirnov Statistic	largest
Rankings	ROC Index (concordance)	largest
	Gini Coefficient	largest
Estimates	Average Squared Error	smallest
	Schwarz's Bayesian Criterion	smallest
	Log-Likelihood	largest

Model Comparison Node



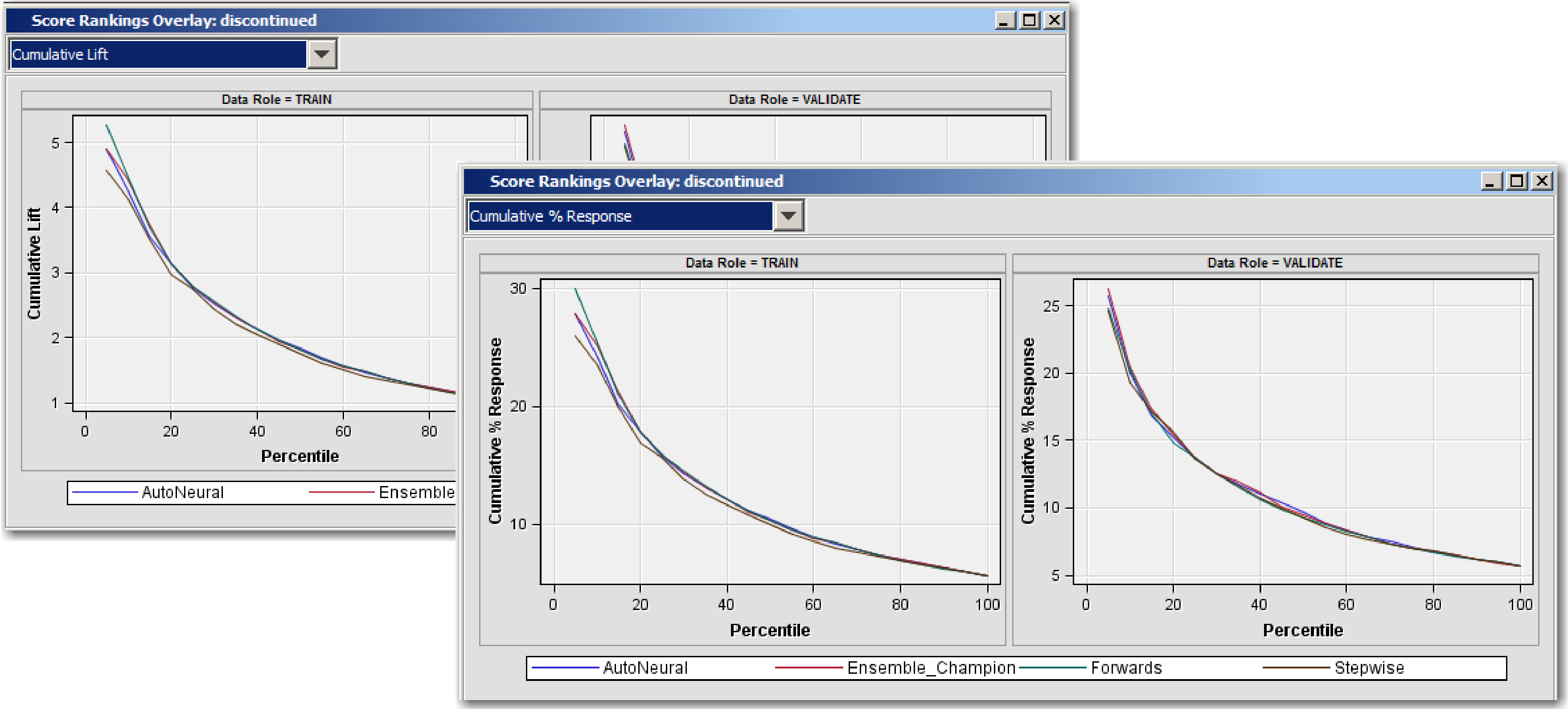
The Model Comparison node provides a common framework for comparing models and predictions from any of the modeling tools (such as Regression, Decision Tree, and Neural Network tools). The comparison is based on standard model fits statistics as well as potential expected and actual profits or losses that would result from implementing the model. The node produces the following charts that help to describe the usefulness of the model: lift, profit, return on investment, receiver operating curves, diagnostic charts, and threshold-based charts.

AIC	Captured Response
ASE	KS Statistic
MSE	Misclassification
ROC	Average Profit/Loss
Gain	Cumulative Lift
Lift	Cumulative Captured Response
Gini	Cumulative Percent Response

Available for training, validation and test datasets

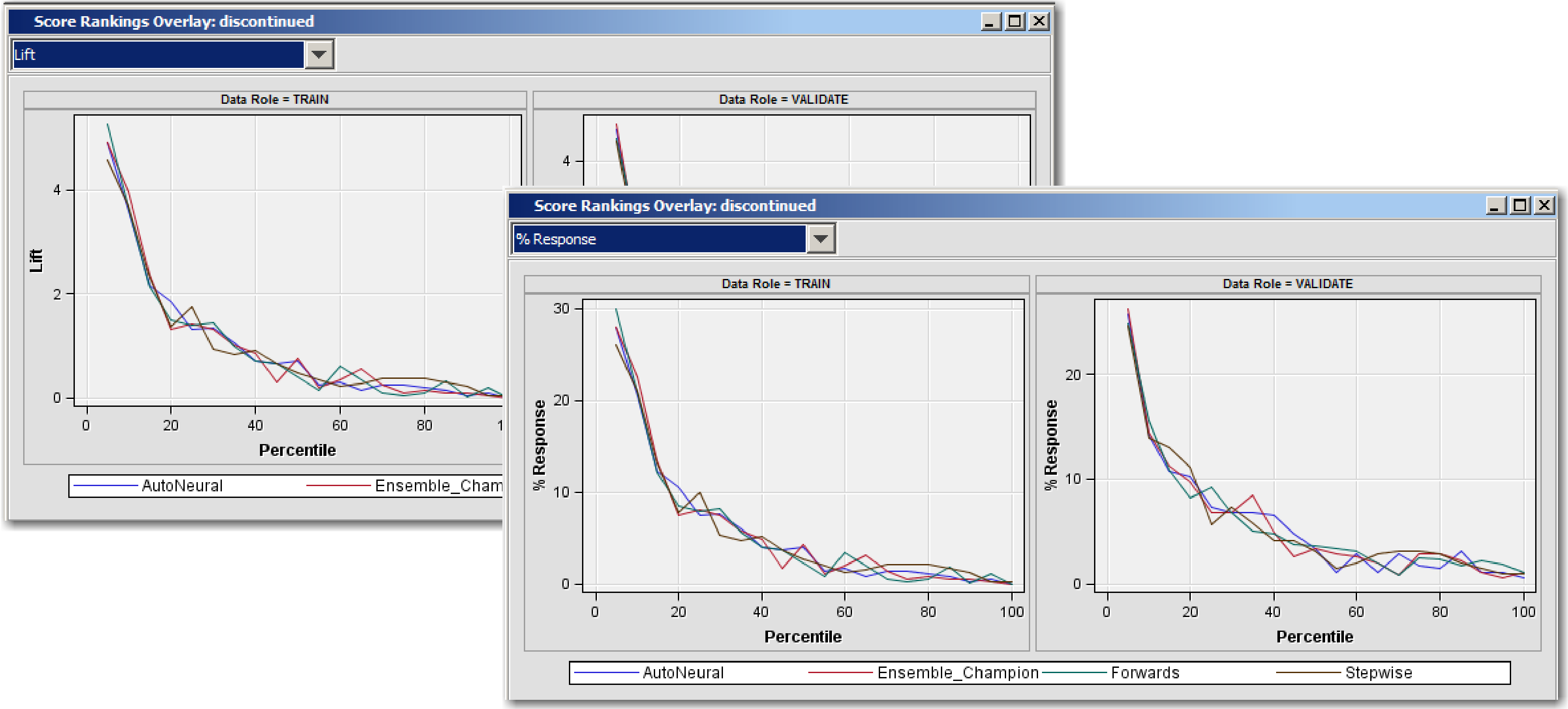
Assess

Cumulative charts



Assess

Non-Cumulative charts



SAS® Enterprise Miner™
Model Comparison Node

Best Model

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate ▲	Train: Misclassification Rate	Valid: Lift	Train: Schwarz's Bayesian Criterion
Y	Reg4	Reg4	Regression DT	TARGET...	Donated ...	0.249441	0.24965	1.539784	15059.33
	HPDMFo...	HPDMFo...	HP Forest	TARGET...	Donated ...	0.249957	0.249797	1.429799	.
	HPReg4	HPReg4	HP Regression stepwise	TARGET...	Donated ...	0.250473	0.249428	1.546658	.
	Reg5	Reg5	Regression PC	TARGET...	Donated ...	0.250645	0.249133	1.443547	14993.11
	HPReg	HPReg	HP Reg - Backward	TARGET...	Donated ...	0.250817	0.249281	1.457295	.
	HPReg3	HPReg3	HP Reg forward	TARGET...	Donated ...	0.250989	0.247585	1.374807	.
	Reg2	Reg2	Regression Forward	TARGET...	Donated ...	0.251161	0.247585	1.361059	15075.82
	Reg3	Reg3	Regression Stepwise	TARGET...	Donated ...	0.251161	0.247585	1.361059	15075.82
	Reg	Reg	Regression Backward	TARGET...	Donated ...	0.251849	0.247732	1.361059	15075.56
	HPReg2	HPReg2	HP Reg Fast Backward	TARGET...	Donated ...	0.252193	0.248838	1.539784	.
	Reg8	Reg8	Regression 2 Poly	TARGET...	Donated ...	0.253226	0.246478	1.484792	15017.38
	Reg6	Reg6	Regression Full	TARGET...	Donated ...	0.253398	0.246773	1.622272	15639.84
	Reg9	Reg9	Reg 2-way Int 2 Poly	TARGET...	Donated ...	0.258214	0.241463	1.429799	16427.17
	Reg7	Reg7	Regression 2-way Interactions	TARGET...	Donated ...	0.295544	0.21211	1.127342	33523.52

SAS Enterprise Miner assumes decision processing and selects the model with the lowest misclassification rate when there is a binary target.

Which?

Model Assessment

Criterion

- Decisions/Assessment
  - Accuracy/Misclassification
  - *Profit/Loss*
  - *Inverse prior threshold*
- Estimates
  - Average squared error
  - SC (SBC or BIC)
- Rankings
  - ROC Index
  - Gini coefficients

Defining Measures of Success for Predictive Models

SAS Enterprise Miner Help under Model Comparison for additional information

sas

Deploying the Model

Best Practices

Model Deployment

- Reporting Results
- Clean up and back up
- Monitor performance

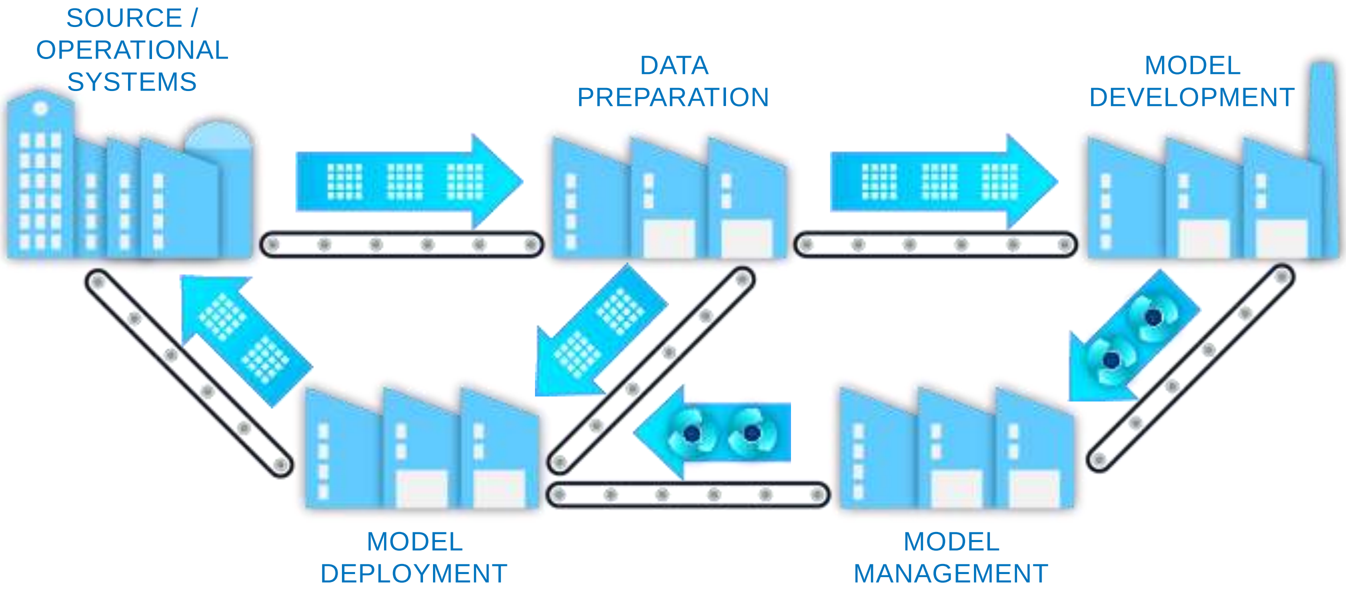
Best Practices

Model Deployment

- Incorporate and share knowledge
- Automate ETL (Extract, Transform, Load)
- Automate process

Ultimate Goal

SAS MODEL FACTORY



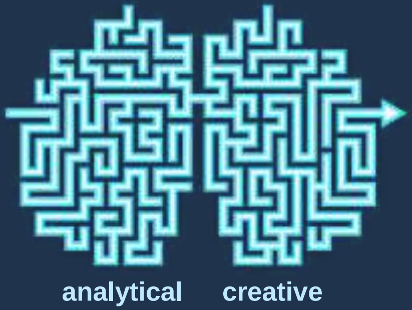
Best Practices

Format of Presentation

- Background & General Guidance
- Developing the Data
- Developing & Delivering the Model

Best Practice

Be analytically savvy and creative



It’s both science *and* art!

Resources

Ready to Get on the Fast Track with Enterprise Miner?

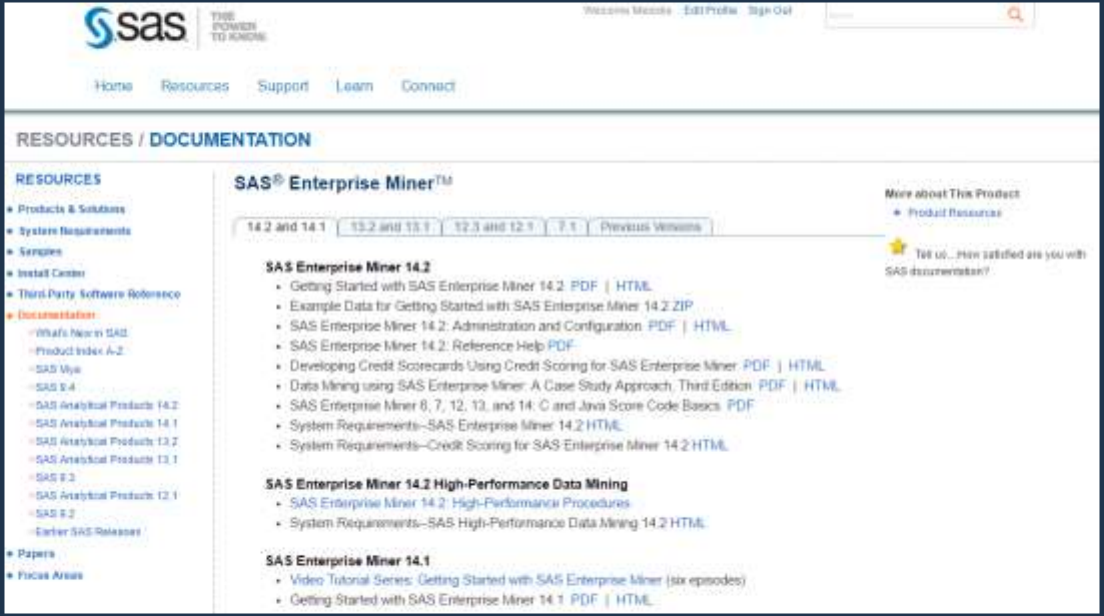
Visit sas.com/learn-em

and sign up to receive EM technical resources, tips & tricks delivered directly from Brett Wujek, Sr. Data Scientist from SAS R&D

SAS® Enterprise Miner™

Getting Started Documentation

- Using same data from "Getting Started with SAS® Enterprise Miner™" documentation
- Both the data and the documentation are available on support.sas.com http://support.sas.com/documentation/onlinedoc/miner/



***Tab and Scroll to find your version*

Further Reading

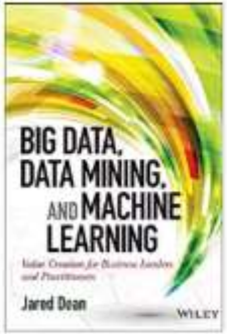
Papers

- Identifying and Overcoming Common Data Mining Mistakes by Doug Wielenga, SAS Institute Inc., Cary, NC
- Best Practices for Managing Predictive Models in a Production Environment by Robert Chu, David Duling, Wayne Thompson , SAS Institute Cary, NC
- From Soup to Nuts: Practices in Data Management for Analytical Performance by David Duling, Howard Plemmons, Nancy Rausch, SAS Institute Cary, NC

- (All available on support.sas.com)

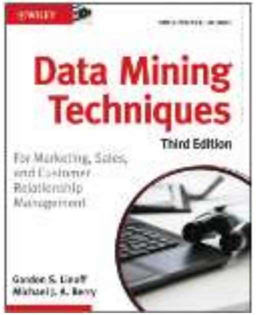
Resources

Suggested Reading



Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners
By Jared Dean

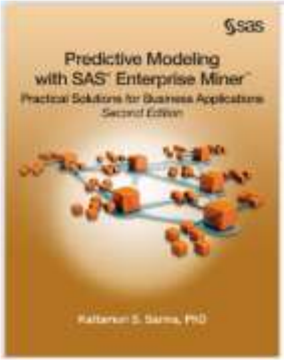
Available on Amazon



Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management
by Gordon S. Linoff and Michael J. A. Berry

Available on Amazon

Resources
Suggested Reading



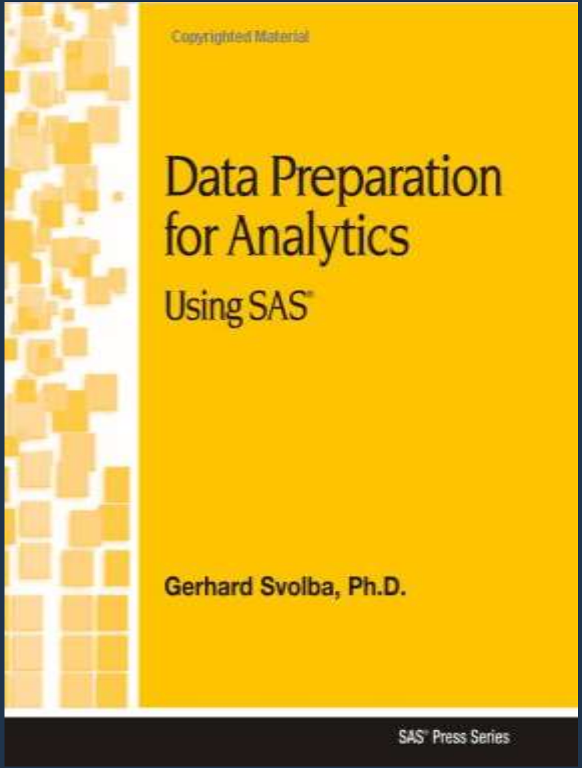
Predictive Modeling with SAS Enterprise Miner: Practical Solutions for Business Applications, Second Edition, Edition 2
By Kattamuri S. Sarma, PhD

Available on Amazon

Applied Analytics Using SAS Enterprise Miner
By: SAS

Available on Amazon

Data Preparation for Analytics Using SAS®



- ISBN: 978-1-59994-047-2
  - SAS Bookstore
  - Amazon
    - Also available for Kindle®
- Author Page
- Example Code and Data


Online. Everyday.
"I always learn something new when I post in this forum. Just what I needed…"
SAS Online Community
Communities.sas.com/data-mining
Support

sas

Questions?

Thank you for your time and attention!

Connect with me:

LinkedIn: https://www.linkedin.com/in/melodierush

Twitter: @Melodie_Rush

sas.com