Proven Practices for Predictive Modeling

Melodie Rush
Customer Success Principal Data Scientist

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

Copyright © SAS Institute Inc. All rights reserved.

Agenda

SAS Enterprise Miner &
SAS Visual Data Mining and Machine Learning

Best practices for creating a predictive model

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

Copyright © SAS Institute Inc. All rights reserved.

Best practices to help you meet and exceed your goals

Faster model development
More *useful* models
Superior models

Copyright © SAS Institute Inc. All rights reserved.

Disclaimers

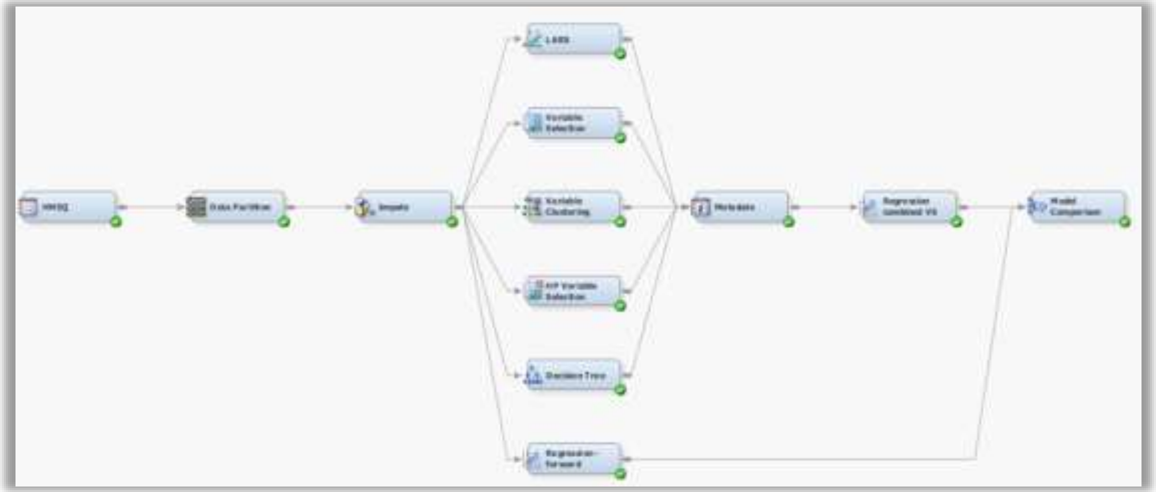
- The choice of “Best Practices” is highly subjective.
- Certain suggested practices may not be suitable for a particular situation.
- It is the responsibility of a predictive modeler to critically evaluate methods and select the best method for a particular situation.
- This presentation represents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

Copyright © SAS Institute Inc. All rights reserved.

SAS® Enterprise Miner™

Overview

Streamlines data mining process and allows you to create accurate predictive and descriptive analytical models in a drag-and-drop GUI.



[More info here]

Copyright © SAS Institute Inc. All rights reserved.

1. Integrated data mining and machine learning
2. Open platform with access from SAS and Open Source
3. Modern machine learning algorithms
4. Cloud-ready and industrial scalable analytics

SAS® Visual Data Mining and Machine Learning is an end-to-end machine learning solution on the most advanced analytics platform.

Copyright © SAS Institute Inc. All rights reserved.

Background

Analytics Cycle and the Modeling Process

Copyright © SAS Institute Inc. All rights reserved.

Why use Predictive Modeling?

To turn increasing amounts of raw data into useful information

Copyright © SAS Institute Inc. All rights reserved.

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

Copyright © SAS Institute Inc. All rights reserved.

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

Copyright © SAS Institute Inc. All rights reserved.

The Goal? Scoring!

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

Copyright © SAS Institute Inc. All rights reserved.

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.

Copyright © SAS Institute Inc. All rights reserved.

Why?

- Consider a group of subjects whose relevant behavior is unknown.
- The <u>same</u> 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**.

Copyright © SAS Institute Inc. All rights reserved.

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.)

Copyright © SAS Institute Inc. All rights reserved.

General Guidance

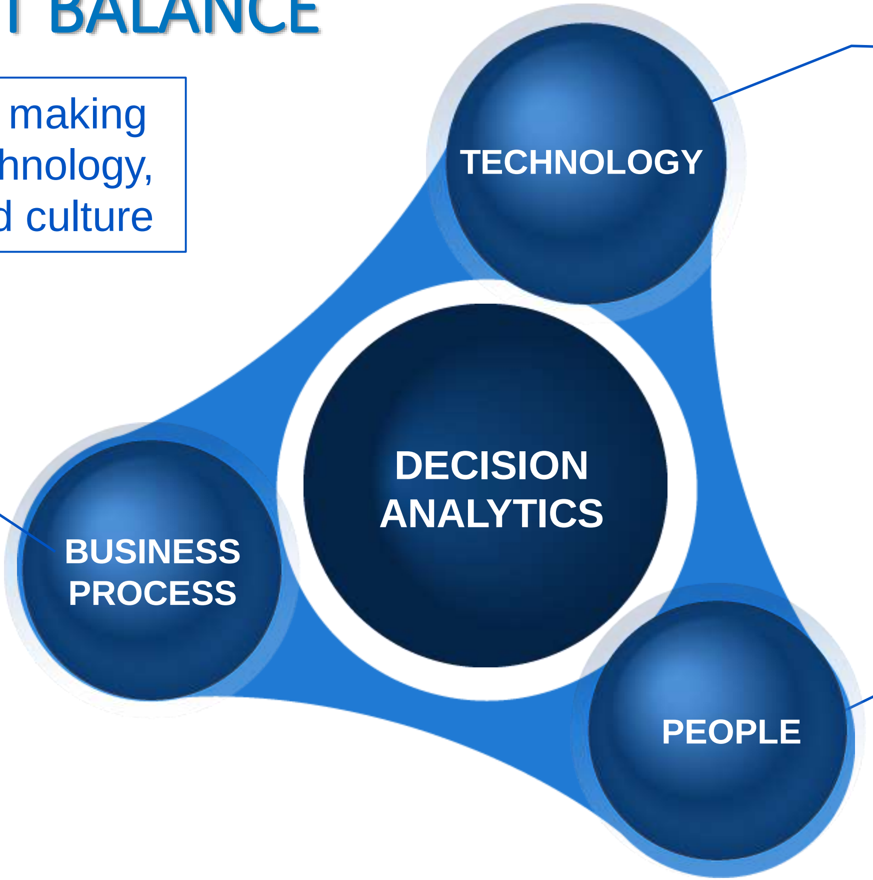
Analytics Cycle and the Modeling Process

Copyright © SAS Institute Inc. All rights reserved.

ITS ALL ABOUT BALANCE

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

- Planning
- Project methodology
- Standards
- Continuous Process Improvement



- Reporting
- Dashboards
- Information management
- Problem-specific business solutions
- Predictive analytics
- Hardware

- Vision & Leadership
- Team composition
- Enterprise authority

Copyright © SAS Institute Inc. All rights reserved.

Lifecycle Best Practice

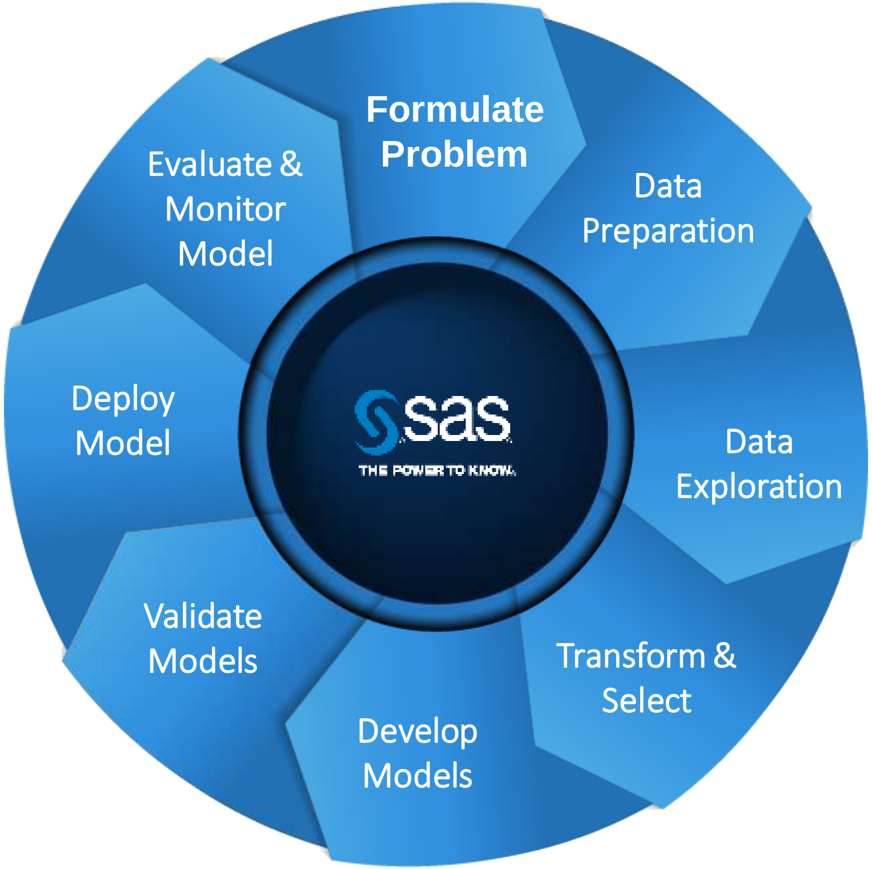
Involve all the relevant people/roles

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

Copyright © SAS Institute Inc. All rights reserved.

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.

Copyright © SAS Institute Inc. All rights reserved.

Best Practice

Begin with the End in Mind

Copyright © SAS Institute Inc. All rights reserved.

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?

Copyright © SAS Institute Inc. All rights reserved.

Best Practices

Business considerations before you model

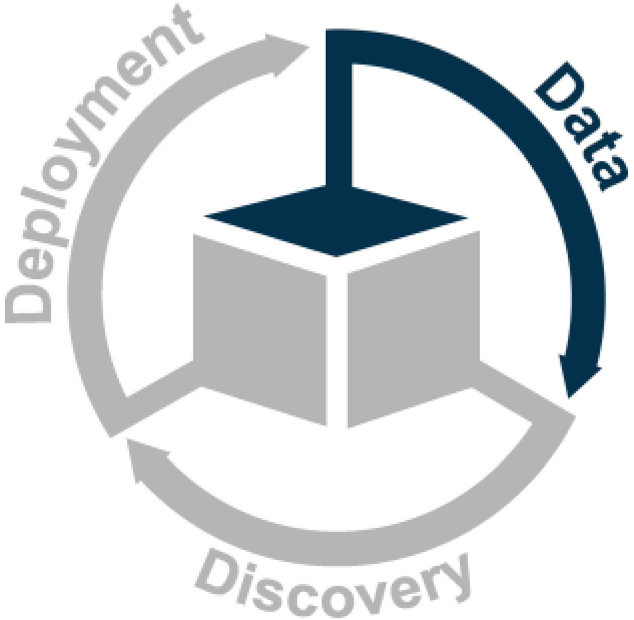
- Thoroughly understand the business/marketing objectives
- Detail the precise (planned) usage for the output
- Define the target variable (the outcome being modeled / predicted)
- Formulate a theoretical model: $Y = f(X_1, X_2, \ldots)$ ← fill-in the likely X's

Copyright © SAS Institute Inc. All rights reserved.

The SAS Platform

Copyright © SAS Institute Inc. All rights reserved.

Essential Data Tasks



- Collect and organize data
- Divide the data
- Address rare events
- Manage missing values
- Add unstructured data
- Extract features
- Handle extreme or unusual values
- Select useful inputs

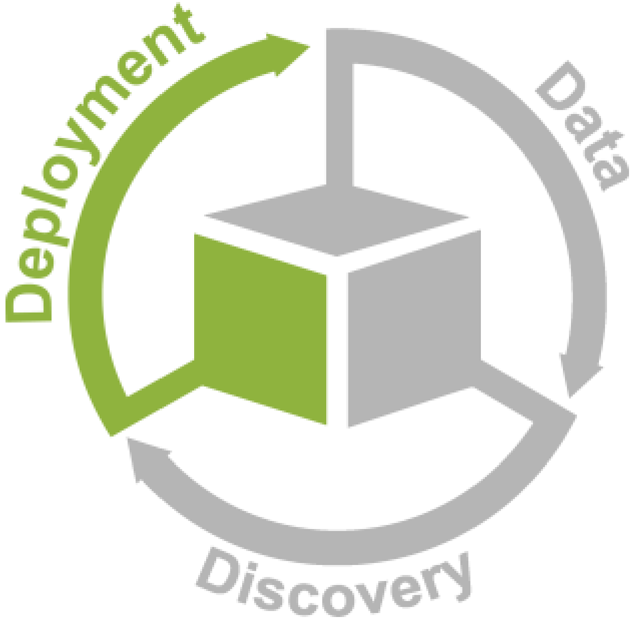
Copyright © SAS Institute Inc. All rights reserved.

Essential Discovery Tasks

- Select an algorithm
- Improve the model
- Optimize complexity of the model
- Regularize and tune hyperparameters of the model
- Build ensemble models
- Attempt other algorithms

Copyright © SAS Institute Inc. All rights reserved.

Essential Deployment Tasks



- Assess models
- Compare models
- Score the champion model
- Monitor model performance over time
- Update the model as needed

Copyright © SAS Institute Inc. All rights reserved.

Developing the Data

Copyright © SAS Institute Inc. All rights reserved.

Best Practices

Optimizing Data

Determining Data

Selecting Target

Preparing Variables

Copyright © SAS Institute Inc. All rights reserved.

Determining Data

Copyright © SAS Institute Inc. All rights reserved.

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

Copyright © SAS Institute Inc. All rights reserved.

Best Practices

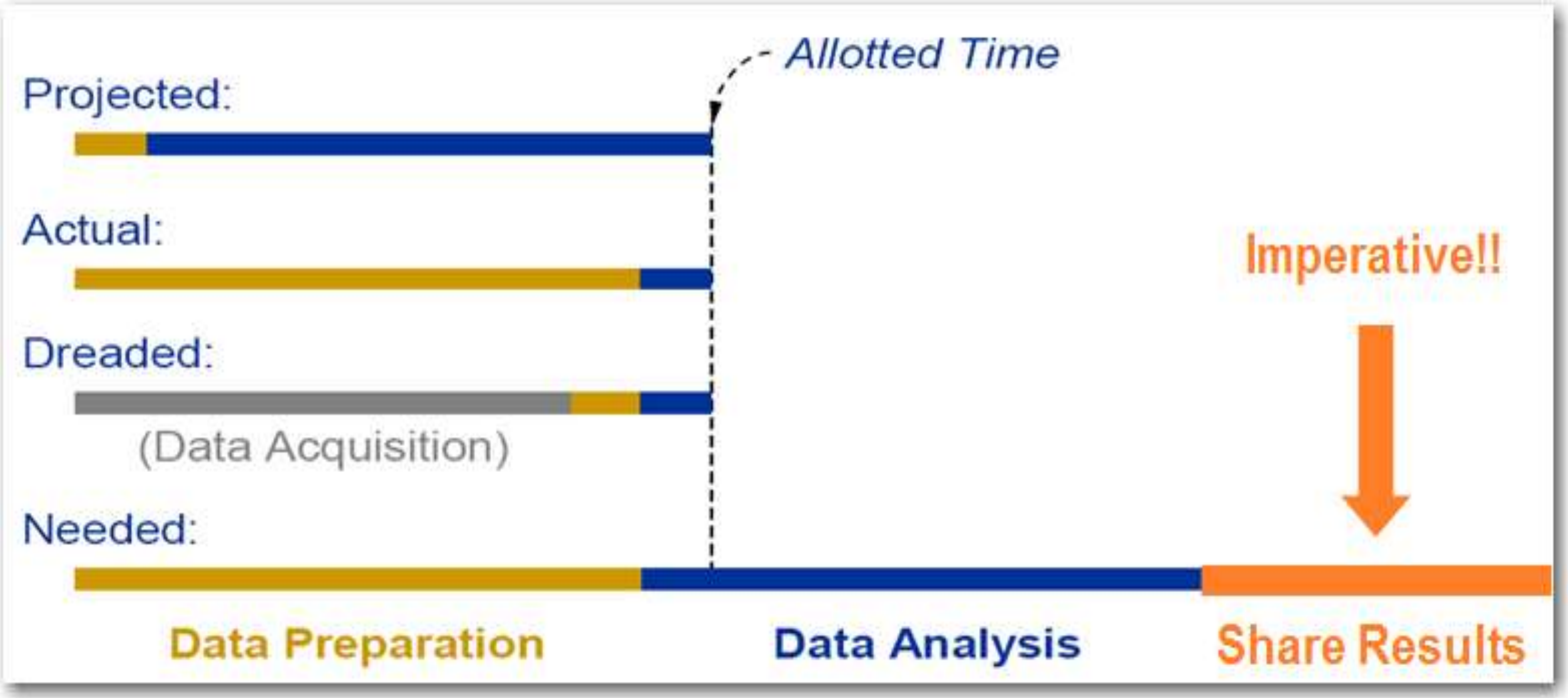
Technical Considerations Before Modeling

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

Copyright © SAS Institute Inc. All rights reserved.

Best Practice

Allow sufficient time for all aspects



Copyright © SAS Institute Inc. All rights reserved.

Sample

Copyright © SAS Institute Inc. All rights reserved.

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.

Copyright © SAS Institute Inc. All rights reserved.

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**.

Copyright © SAS Institute Inc. All rights reserved.

Partitioning

Honest Assessment

Copyright © SAS Institute Inc. All rights reserved.

SAMPLE

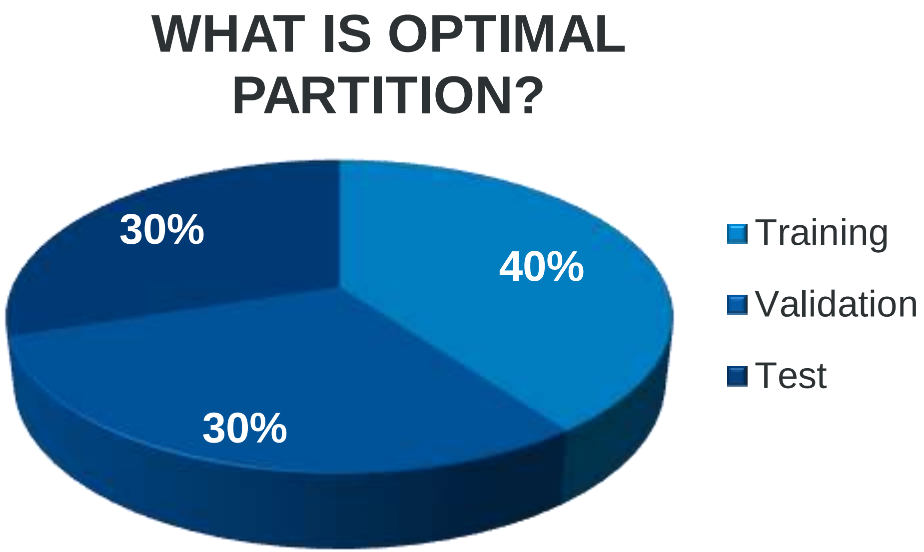
Data Partitioning

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

Copyright © SAS Institute Inc. All rights reserved.

Sample

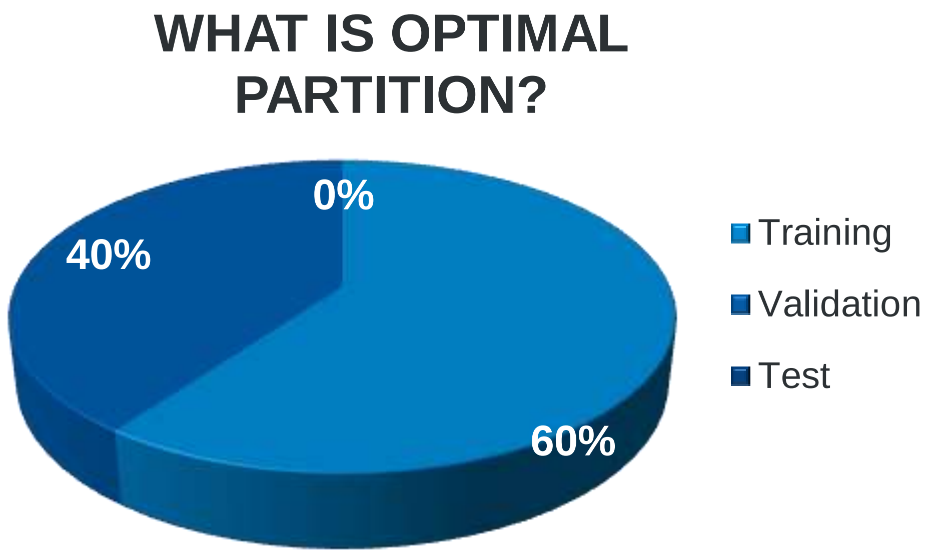
SAMPLE: Data Partitioning



Copyright © SAS Institute Inc. All rights reserved.

Best Practice

SAMPLE: Data Partitioning

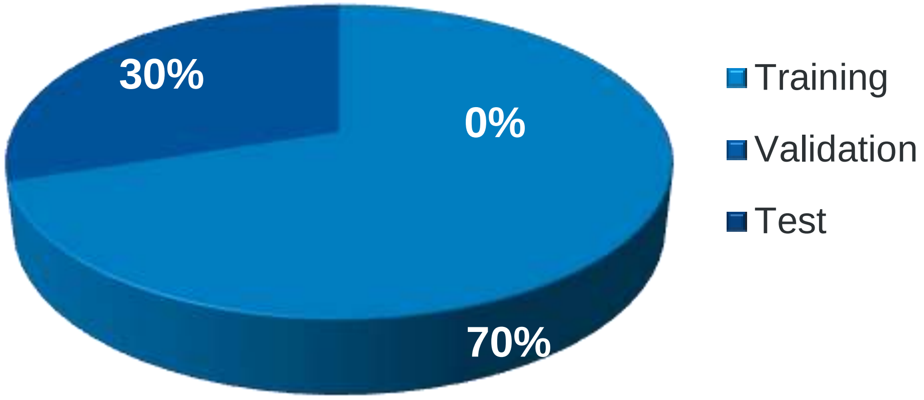


Copyright © SAS Institute Inc. All rights reserved.

Best Practice

Sample: Data Partitioning

WHAT IS OPTIMAL PARTITION?



It depends!

Copyright © SAS Institute Inc. All rights reserved.

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?

Copyright © SAS Institute Inc. All rights reserved.

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.

Copyright © SAS Institute Inc. All rights reserved.

Selecting Target

Copyright © SAS Institute Inc. All rights reserved.

Choosing your target

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

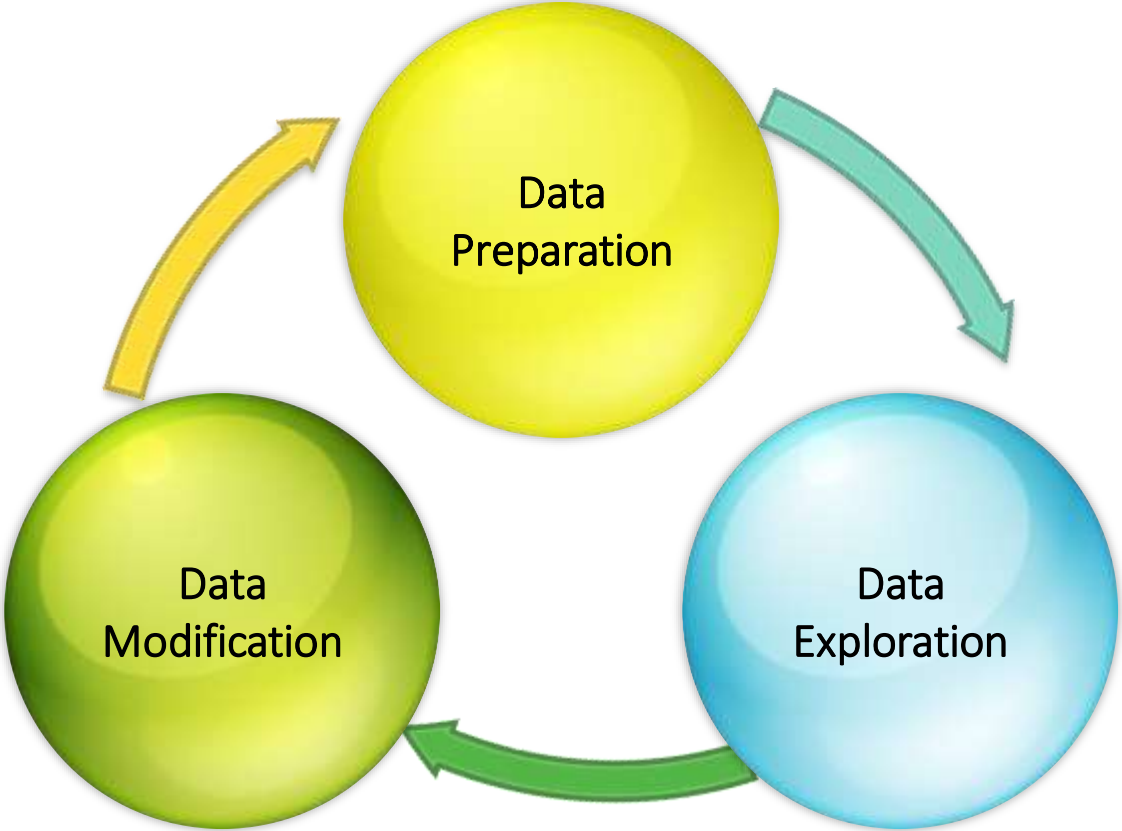
Copyright © SAS Institute Inc. All rights reserved.

Preparing Data

Copyright © SAS Institute Inc. All rights reserved.

EXPLORE & MODIFY

Iterative Relationship with Data Preparation



Copyright © SAS Institute Inc. All rights reserved.

Best Practice

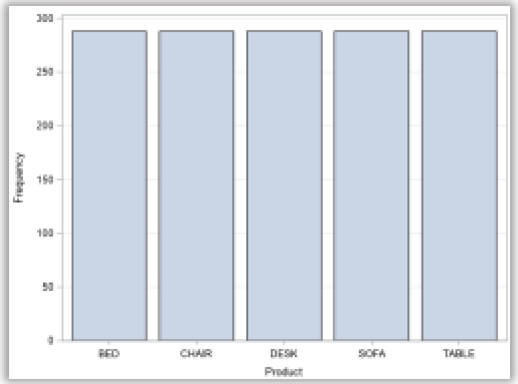
Explore & Modify: Getting the Most out of Data

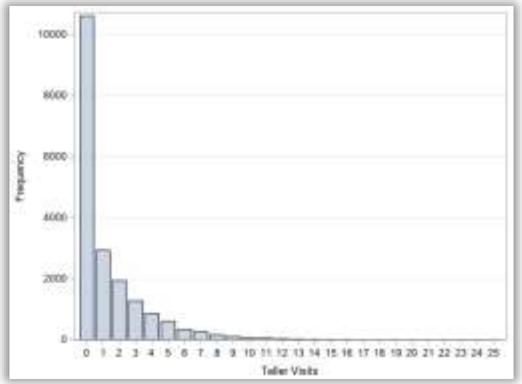
- Once you have an analytics-ready table:
  - Examine *Categorical* Variables
  - Examine *Continuous* Variables
  - Explore *Missing* Values
  - *Cluster* Variables

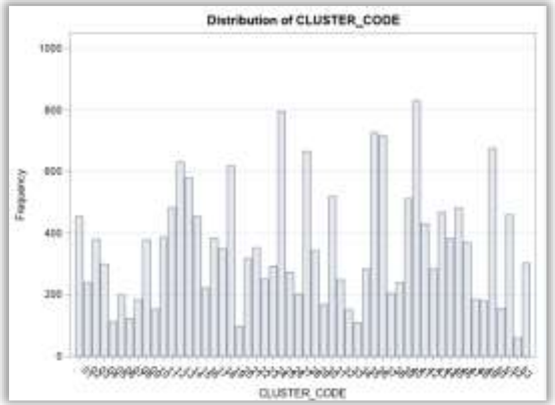
Copyright © SAS Institute Inc. All rights reserved.

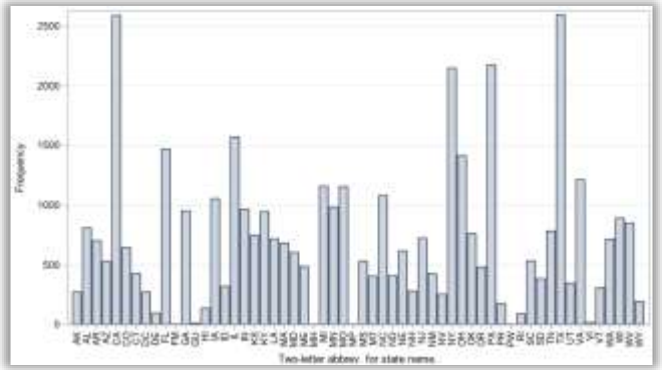
Explore & Modify

Categorical Variables





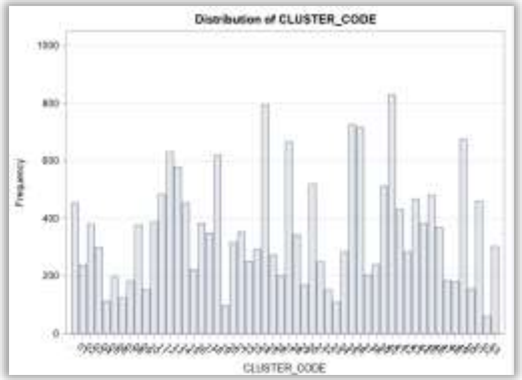


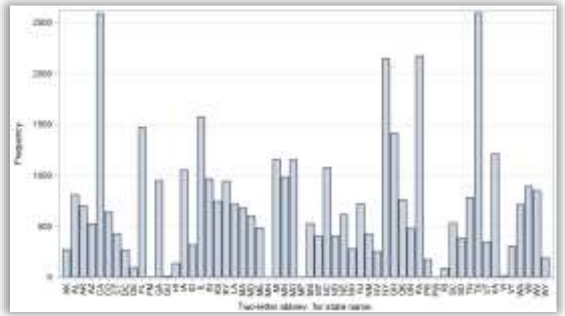


Copyright © SAS Institute Inc. All rights reserved.

Explore & Modify

Categorical Variables





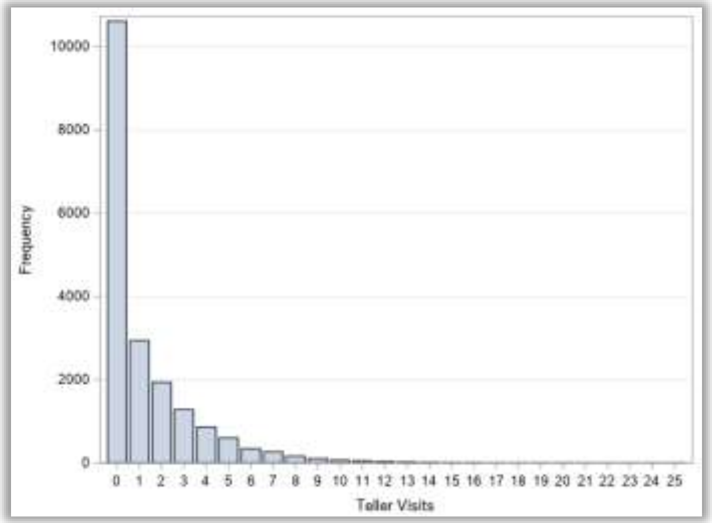
Too many overall values

- 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

Copyright © SAS Institute Inc. All rights reserved.

Explore & Modify
Categorical Variables

Levels that rarely occur

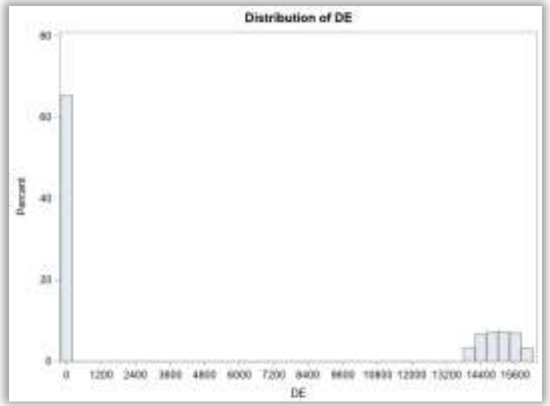


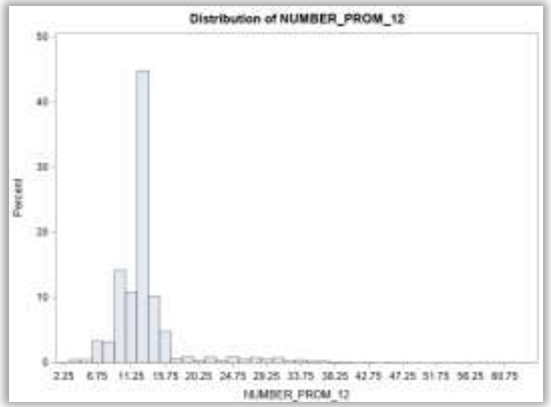
- Group infrequently occurring values together as "other"
- Judiciously combine a less frequently occurring level with a more frequent one where it makes business sense
- Consider a less granular level of a hierarchy

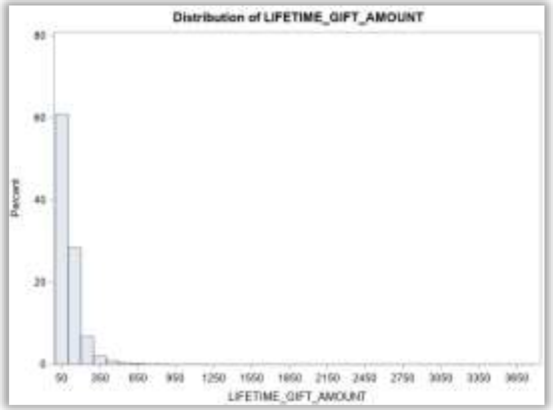
Copyright © SAS Institute Inc. All rights reserved.

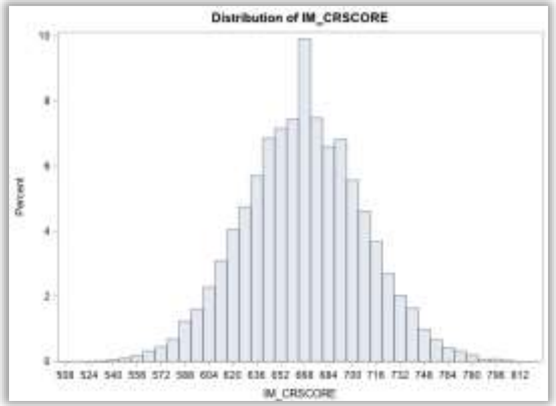
Explore & Modify

Continuous Variables





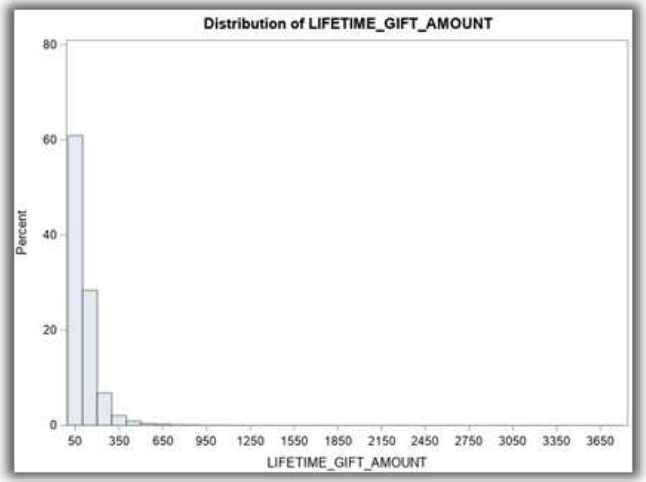




Copyright © SAS Institute Inc. All rights reserved.

Explore & Modify

Continuous Variables



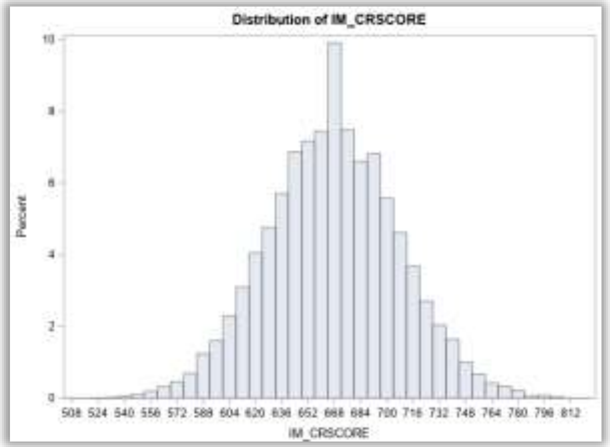
Extremely skewed predictors

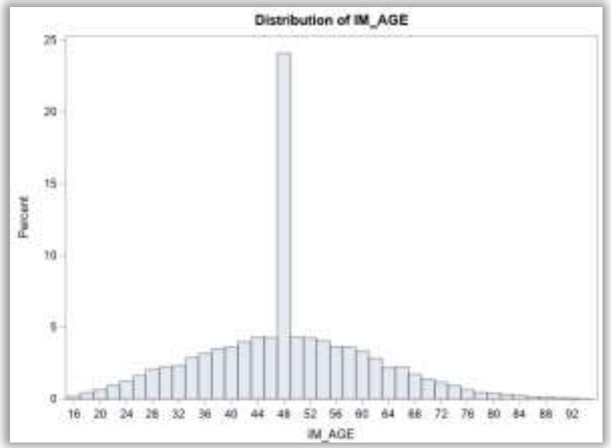
- Consider transformations that stabilize variance and generate more support across the range of values
- Consider binning transformation with appropriate number of bins to enable each portion of the ranges to be weighed appropriately

Copyright © SAS Institute Inc. All rights reserved.

Explore & Modify

Continuous Variables





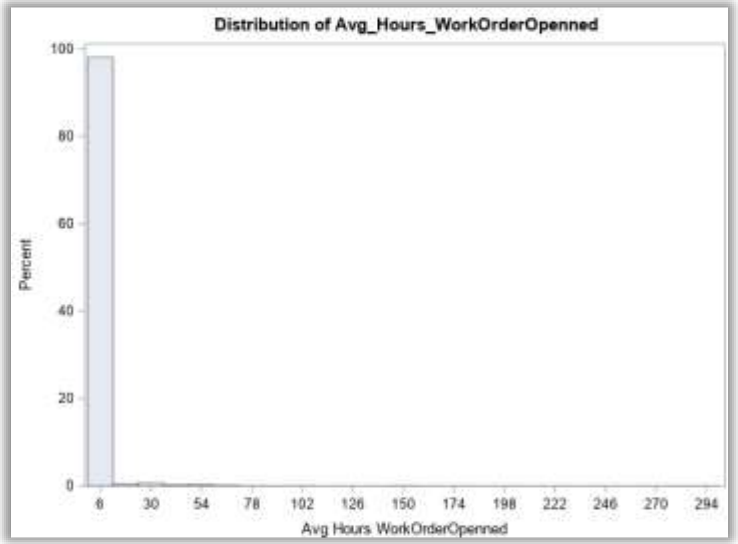
Spike and a Distribution

- 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

Copyright © SAS Institute Inc. All rights reserved.

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

Copyright © SAS Institute Inc. All rights reserved.

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*

Copyright © SAS Institute Inc. All rights reserved.

Explore & Modify

Variables for Clustering

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

Copyright © SAS Institute Inc. All rights reserved.

Selecting Variables

Copyright © SAS Institute Inc. All rights reserved.

What?

Variable Selection or Variable Reduction

Variable selection is used to find a subset of the available inputs that accurately predict the output.

Copyright © SAS Institute Inc. All rights reserved.

Why Variable Selection?

- Smaller Data
  - Speed/Performance
    - Decreased Computation Time
    - Decreased Scoring Effort
  - Cost
    - Data Collection
    - Data Cleaning

- Other Statistical Reasons
  - Interpretability
  - Multicollinearity & Irrational Coefficients
  - Missing Data
  - Redundancy
  - Predictive Power
    - Destabilize the parameter estimates
    - Increase the risk of over fitting
    - Noise

The principle of Occam's Razor states that among several plausible explanations for a phenomenon, the simplest is best.

Copyright © SAS Institute Inc. All rights reserved.

Variable Selection
Concepts

Variable Selection

- Regression based
- Criterion Based
- Variable Screening
- Variable Clustering

Variable Combination

- Principal Components uncorrelated linear combinations of *all* input variables

Copyright © SAS Institute Inc. All rights reserved.

SAS® Enterprise Miner™

Methods available

- Regression
- Decision Trees
  - Random Forest
- Variable Selection
  - Stat Explore (one level tree)
  - Variable Selection (Chi-Square & R-Square)
  - LARS/LASSO
  - High Performance Variable Selection
- Variable Clustering
- Principle Components
- Weight of Evidence (WOE)

Copyright © SAS Institute Inc. All rights reserved.

SAS® Visual Data Mining and Machine Learning

Methods available

- Variable Selection
  - Unsupervised Selection (No Target)
  - Supervised Selection
    - Fast Supervised Selection
    - Linear Regression Selection
    - Decision Tree Selection
    - Forest Selection
    - Gradient Boosting Selection
- Variable Clustering

Copyright © SAS Institute Inc. All rights reserved.

Summary of Data Preparation

Topic	Common Challenges	Suggested Best Practices
Data Collection	• Biased data • Incomplete data • High-dimensional data • Sparsity	• Take time to understand the business problem and its context • Enrich the data • Dimension reduction (Feature Extraction, Variable Clustering, and Variable Selection) • Change representation of data (Transformations node)

Copyright © SAS Institute Inc. All rights reserved.

Summary of Data Preparation

Topic	Common Challenges	Suggested Best Practices
Messy Data	• Value ranges as columns • Multiple variables in the same column • Variables in both rows and columns	• Transform the data with SAS code (Code node)
Outliers	• Out-of-range numeric values and unknown categorical values in score data	• Discretization (Transformations node) • Winsorizing (Imputation node)

Copyright © SAS Institute Inc. All rights reserved.

Summary of Data Preparation

Topic	Common Challenges	Suggested Best Practices
Sparse target variables	• Low primary event occurrence rate • Overwhelming preponderance of zero or missing values in target	• Proportional oversampling
Variables of disparate magnitudes	• Misleading variable importance • Distance measure imbalance • Gradient dominance	• Standardization (Transformations node)

Copyright © SAS Institute Inc. All rights reserved.

Summary of Data Preparation

Topic	Common Challenges	Suggested Best Practices
High-cardinality variables	• Overfitting • Unknown categorical values in holdout data	• Binning (Transformations node) • Replacement (Replacement node)
Missing Data	• Information loss • Bias	• Binning (Transformations node) • Imputation (Imputation node)
Strong multicollinearity	• Unstable parameter estimates	• Dimension reduction (Feature Extraction, Variable Clustering, and Variable Selection nodes)

Copyright © SAS Institute Inc. All rights reserved.

Best Practices

Optimizing Data

Determining Data
Selecting Target
Preparing Variables

Copyright © SAS Institute Inc. All rights reserved.

Developing & Delivering the Model

Copyright © SAS Institute Inc. All rights reserved.

Model & Assess

Delivering the Model

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

Copyright © SAS Institute Inc. All rights reserved.

Developing the Model

Copyright © SAS Institute Inc. All rights reserved.

MODEL

Model Development

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

Copyright © SAS Institute Inc. All rights reserved.

BEST PRACTICE

Model Development

Try various techniques and combinations of techniques.

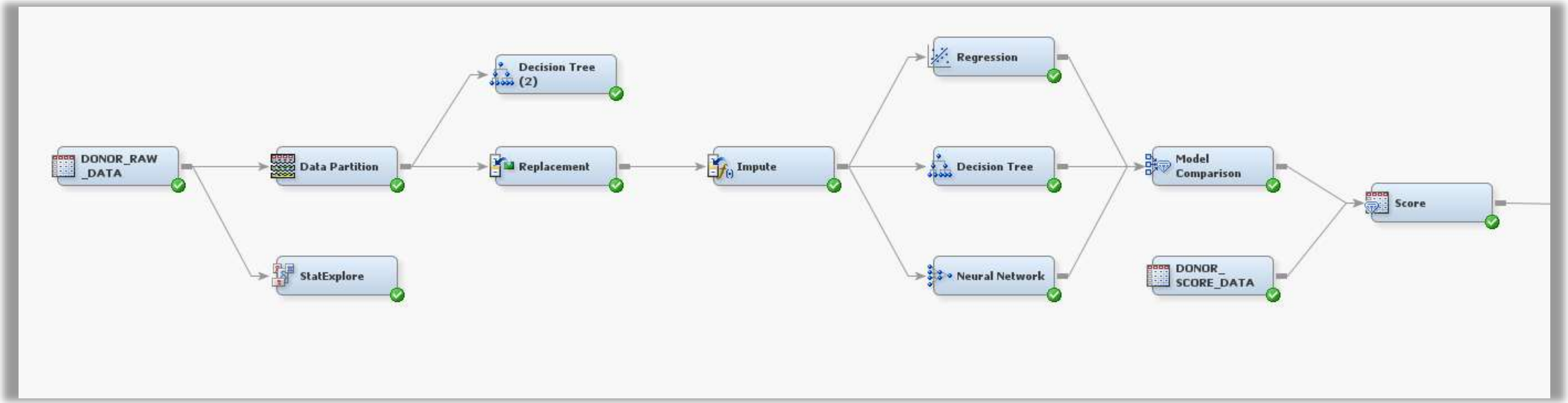
Copyright © SAS Institute Inc. All rights reserved.

Complete List of SAS Enterprise Miner nodes

SAMPLE	Append; Data Partition; File Import; Filter; Merge; Sample; Input Data
EXPLORE	Association; Cluster; Graph Explore; Variable Clustering; DMDB; MultiPlot; Market Basket; StatExplore; Link Analysis; Path Analysis; Variable Selection; SOM/Kohonen
MODIFY	Drop; Impute; Interactive Binning; Principal Components; Replacement; Rules Builder; Transform Variables
MODEL	Decision Tree; AutoNeural; Regression; Neural Network; Partial Least Squares; Dmine Regression; DM Neural; Ensemble; Rule Induction; Gradient Boosting; LARS; MBR; Two Stage; Model Import
	Incremental Response; Survival Analysis; Credit Scoring*; TS Correlation; TS Data Prep; TS Dimension Reduction; TS Decomp.; TS Similarity; TS Exponential Smoothing
	HP Explore; HP Bayesian Network; HP Regression; HP Transform; HP Impute; HP Variable Selection; HP Neural; HP Forest; HP Decision Tree; HP Data Partition; HP GLM; HP SVM; HP Cluster; HP Principal Components
ASSESS	Cutoff; Decisions; Model Comparison; Score; Segment Profile
UTILITY	Control Point; End Groups; Start Groups; Open Source Integration; Reporter; Score Code Export; Metadata; SAS Code; Ext Demo; Save Data; Register Metadata; SAS Viya Code

Copyright © SAS Institute Inc. All rights reserved.

SAS Enterprise Miner



Copyright © SAS Institute Inc. All rights reserved.

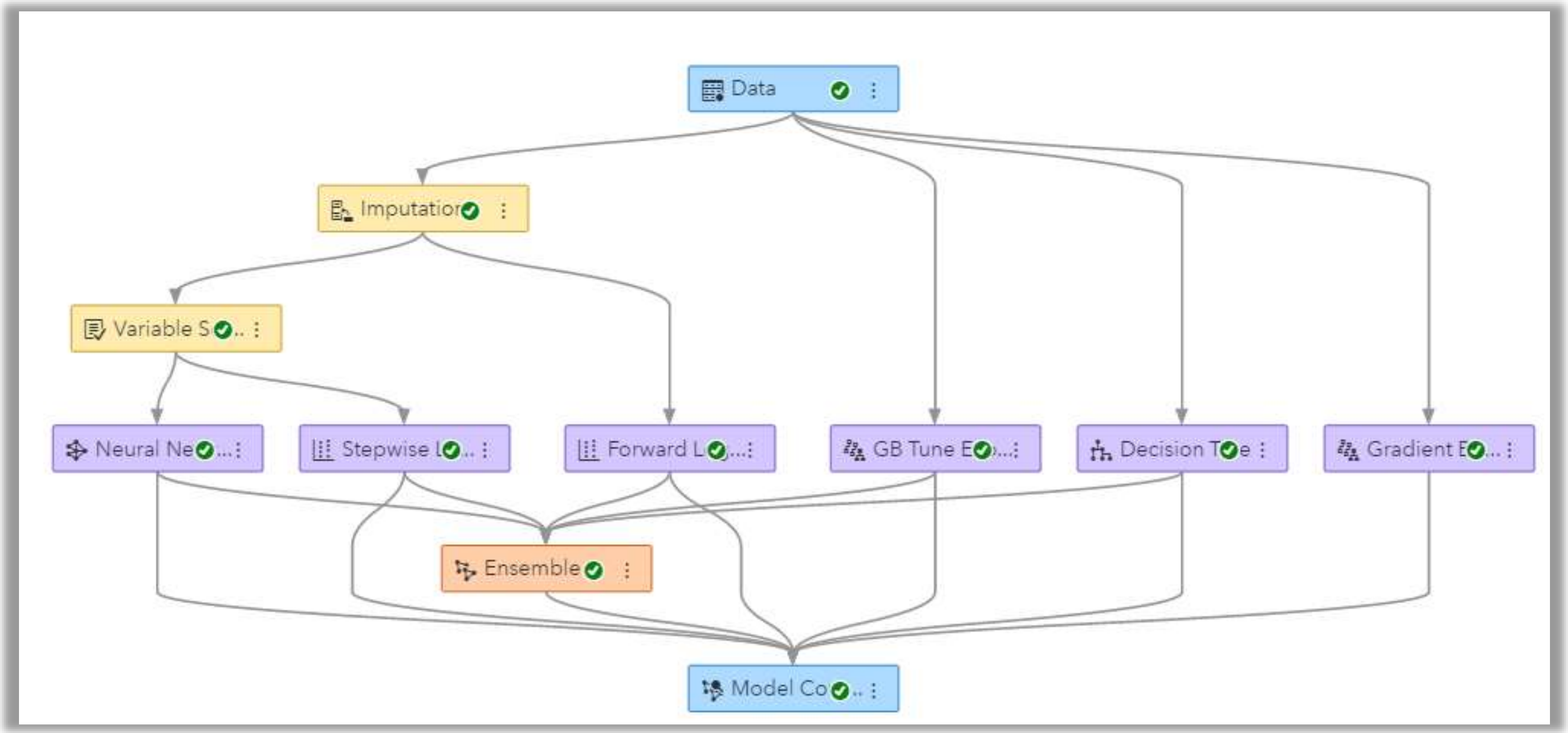
SAS® Visual Data Mining and Machine Learning Capabilities

Analytics	Statistics	Machine Learning	
Analytic Astore Scoring	Cox Proportional Hazards	Automated ML	Model Interpretability
Data Step	Decision Trees	*Data Science Pilot*	*LIME, ICE, PD, Shapley*
Data Transpose	Design Matrix	*Model Composer*	Moving Windows PCA
DS2	General Additive Models	Audio Data Processing	Multitask Learning
Function Compiler	Generalized Linear Models	Bayesian Networks	Network Analytics
FedSQL	Independent Component Analysis	Boolean Rules	Neural Networks
Frequency / Crosstab	Clustering	Cross Validation ML	Random and Isolation Forests
Imputation	*K-means and K-modes*	Deep Learning	Recommendation Engine
Model Assessment	Linear Regression	*Convolutional NN*	Robust PCA
Sampling and Partitioning	Linear Mixed Models	*Deep Forward NN*	Semi-supervised Learning
Sentiment Analysis	Logistic Regression	*Recurrent NN*	Sparse Machine Learning
Sequencing/Pathing Analysis	Model-Based Clustering	*Transfer Learning*	Support Vector Data Description
Text Mining	Model Scoring	Factorization Machine	Support Vector Machine
Variable Binning	Nonlinear Regression	Frequent Item Set Mining	*t*-distributed SNE
Variable Cardinality Analysis	Ordinary Least Squares Regression	Gaussian Mixture Model	Text Parsing
Variable Clustering	Partial Least Squares Regression	Gaussian Process Regression	
Variable Selection	Pearson Correlation	Gradient Boosting	
Variable Summary	Principal Component Analysis	Hyperparameter Auto-tuning	
	Kernel Principal Comp. Analysis	K Nearest Neighbor	
	Quantile Regression	Image Processing (*incl.* Biomedical)	
	Shewhart Control Chart Analysis	Market Basket Analysis	

Last updated 12/10/2019
For documentation please click here

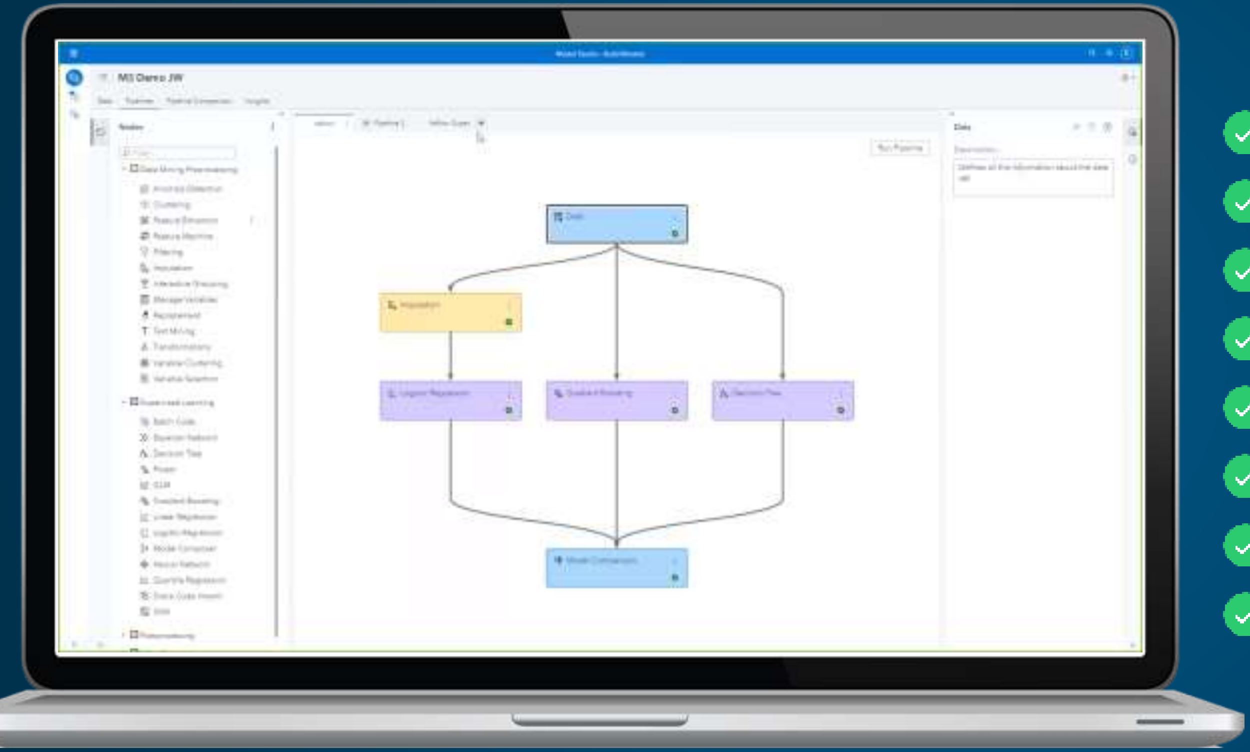
Copyright © SAS Institute Inc. All rights reserved.

SAS Visual Data Mining and Machine Learning



Copyright © SAS Institute Inc. All rights reserved.

Automated Pipelines



- Repository of best practice pipelines
- Models by SAS or by end-user
- Dynamically reads thru data
- Fixes data quality issues w/ ML
- Performs Data transformations
- Recommends & builds models
- Optimizes across models
- Fully editable, no black-box

Copyright © SAS Institute Inc. All rights reserved.

Choosing a Model

Copyright © SAS Institute Inc. All rights reserved.

Best Practices

Model Selection

- Evaluate model metrics
- Consider business knowledge
- Recognize constraints

Copyright © SAS Institute Inc. All rights reserved.

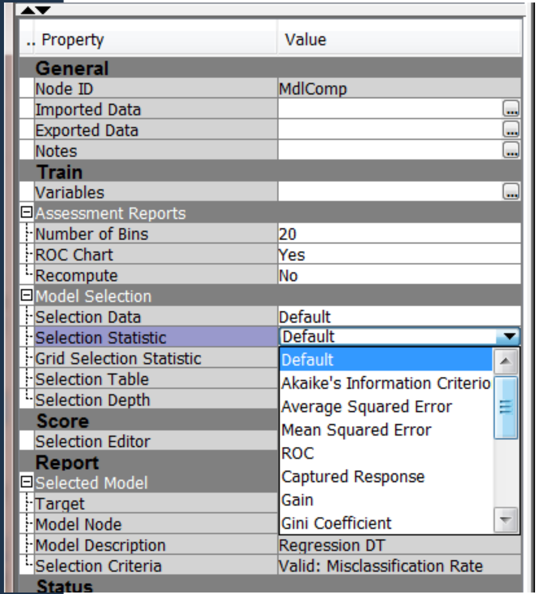
How do we choose?

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

Copyright © SAS Institute Inc. All rights reserved.

SAS® Enterprise Miner™

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.

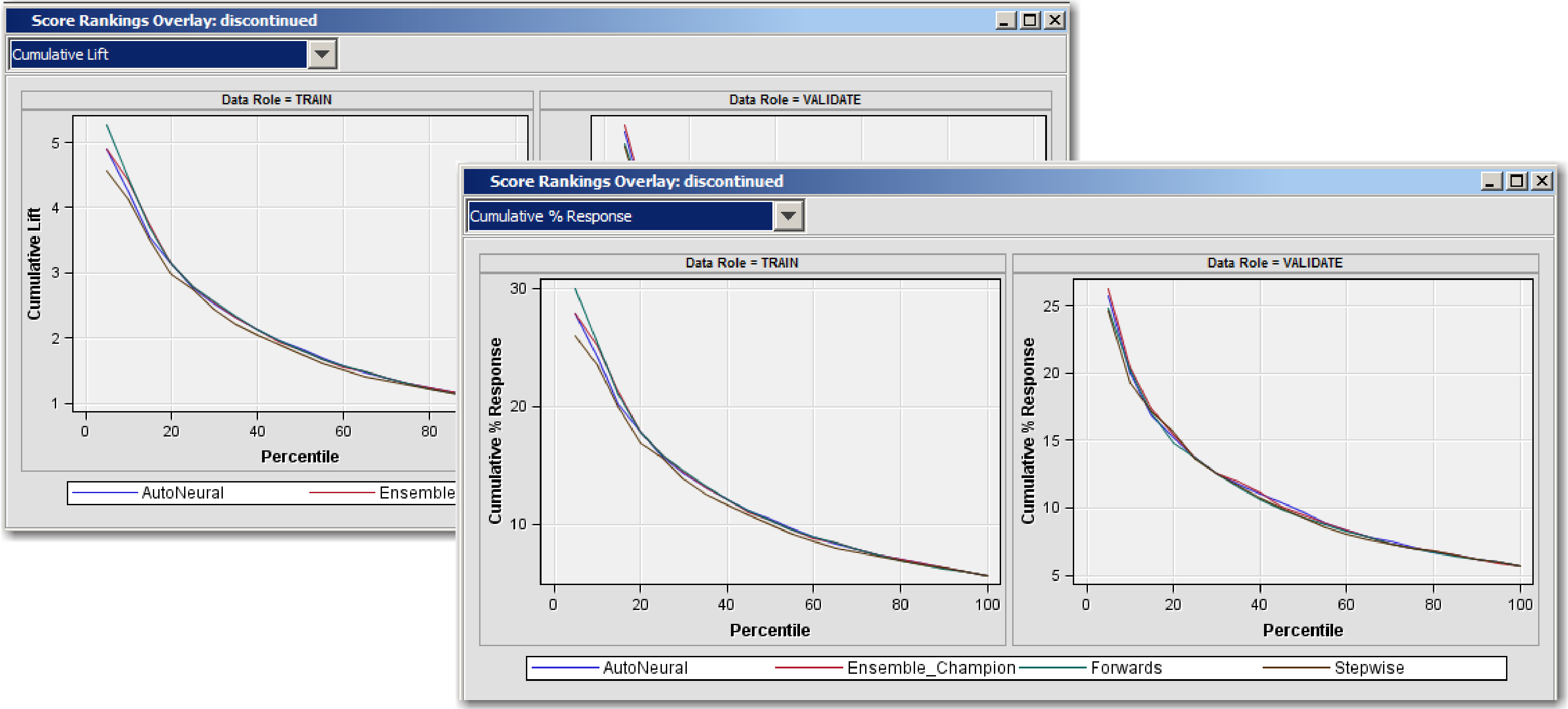
AIC	Captured Response
ASE	KS Statistic
MSE	Misclassification
ROC	Average Profit/Loss
Gain	Cumulative Lift
Lift	Cumulative Captured Response
Gini	Cumulative Percent Response

Available for training, validation and test datasets

Copyright © SAS Institute Inc. All rights reserved.

Assess

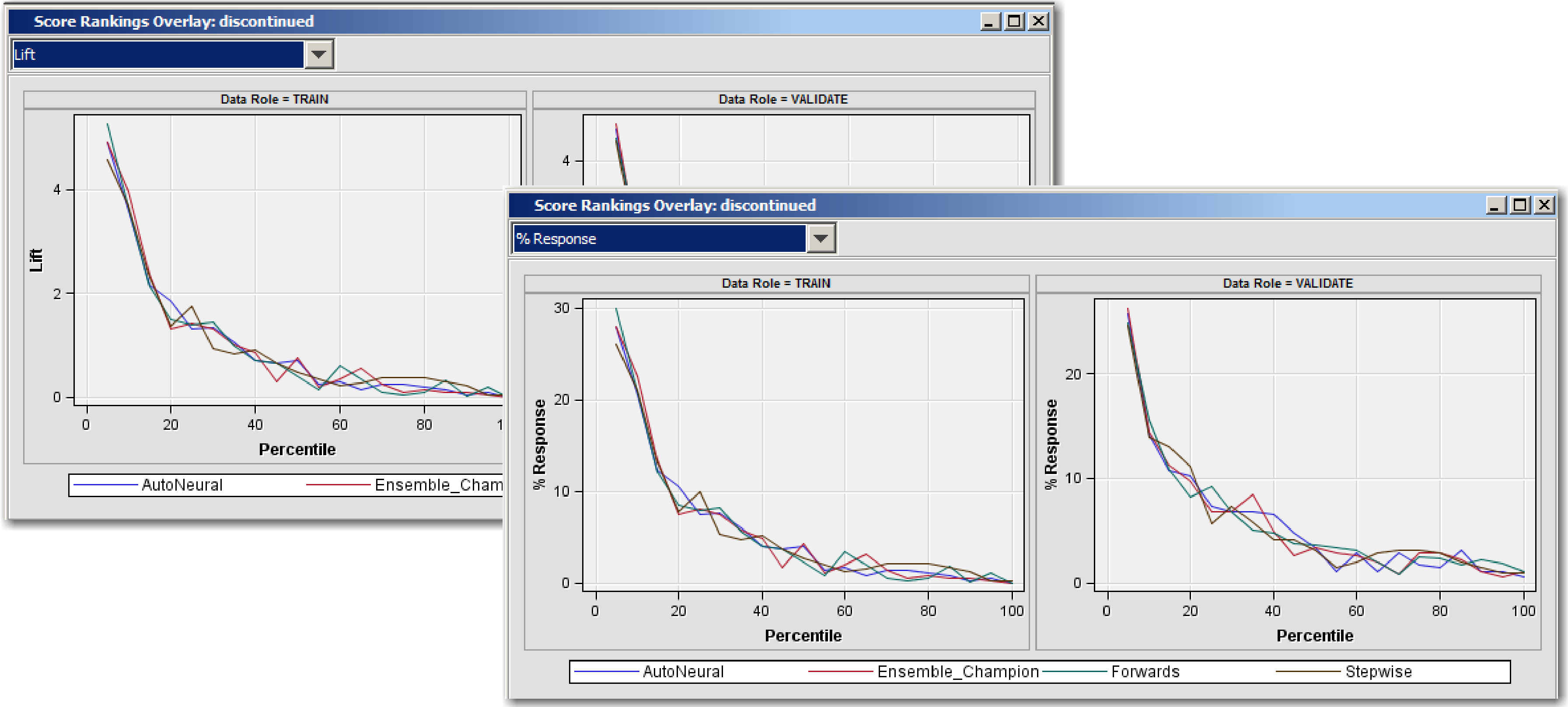
Cumulative charts



Copyright © SAS Institute Inc. All rights reserved.

Assess

Non-Cumulative charts



Copyright © SAS Institute Inc. All rights reserved.

SAS® Enterprise Miner™
Model Comparison Node

Best Model

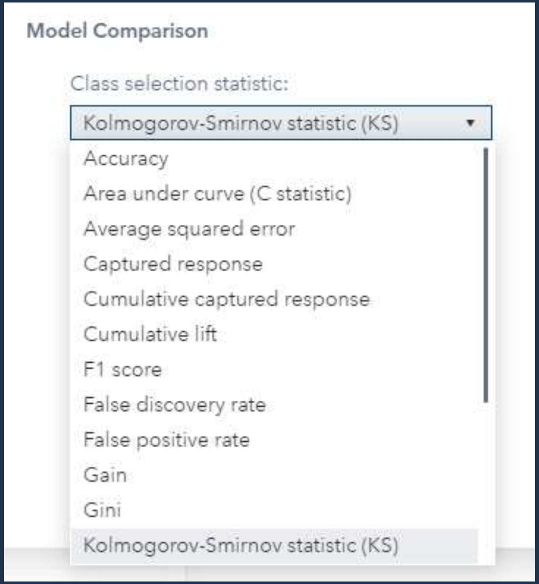
Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate ▲	Train: Misclassification Rate	Valid: Lift	Train: Schwarz's Bayesian Criterion
Y	Reg4	Reg4	Regression DT	TARGET...	Donated ...	0.249441	0.24965	1.539784	15059.33
	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.

Copyright © SAS Institute Inc. All rights reserved.

SAS® Visual Data Mining and Machine Learning™

Model Comparison Node



Accuracy
C statistic
Average Squared Error
Captured response
Cumulative captured response
Cumulative lift
Root average squared error
False Discover rate
False positive rate

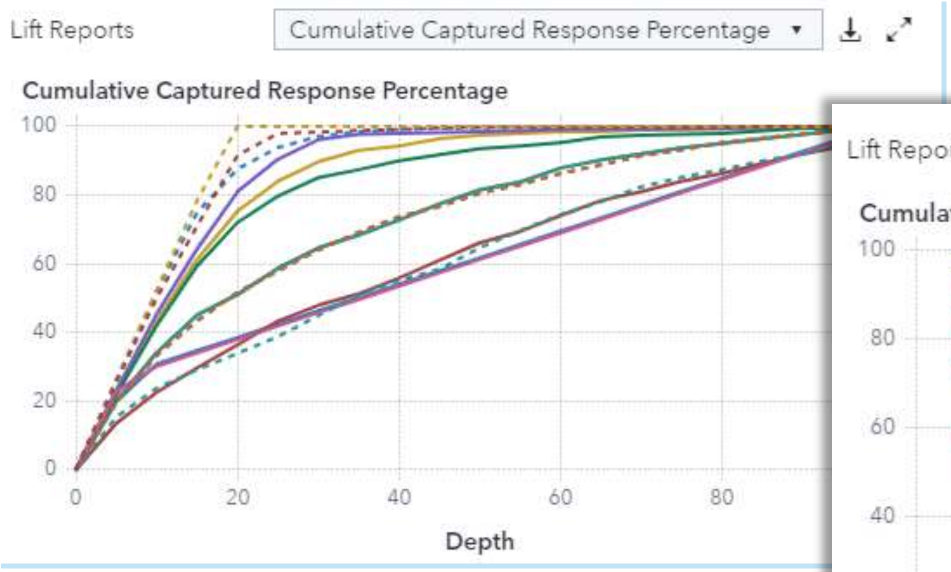
Gain
Gini
KS Statistic
Lift
Misclassification
Multiclass log loss
ROC separation
F1 Score

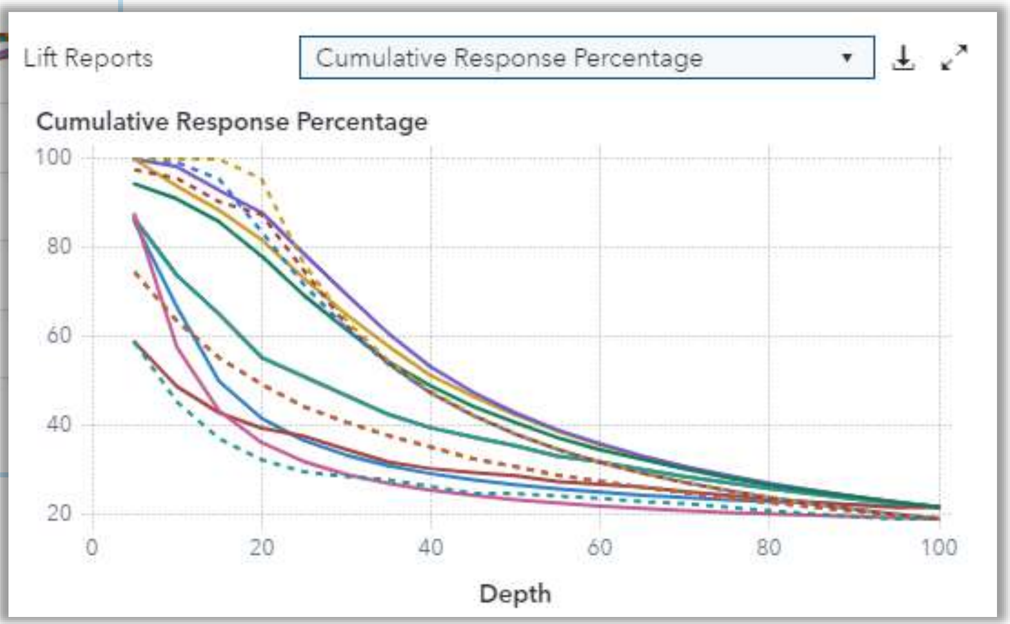
Available for training, validation and test datasets

Copyright © SAS Institute Inc. All rights reserved.

Assess

Cumulative charts

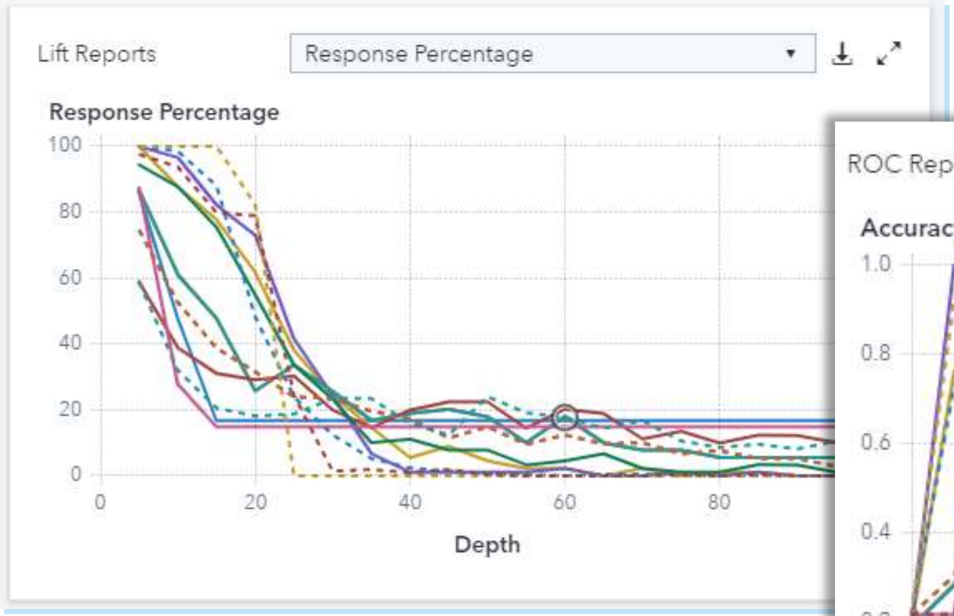




Copyright © SAS Institute Inc. All rights reserved.

Assess

Non-cumulative charts



Copyright © SAS Institute Inc. All rights reserved.

SAS® Visual Data Mining and Machine Learning
Model Comparison Node

Home Loan Default Demo > "Model Comparison" Results

Close

Node Assessment

Model Comparison

Champion	Name	Algorithm Name	Misclassification Rate (Event)	Misclassification Rate
★	GB Tune Explain	Gradient Boosting	0.0699	0.0699
	Gradient Boosting	Gradient Boosting	0.0912	0.0912
	Ensemble	Ensemble	0.1549	0.1549
	Forward Logistic Regression	Logistic Regression	0.1711	0.1711
	Stepwise Logistic Regression	Logistic Regression	0.1711	0.1711
	Decision Tree	Decision Tree	0.1784	0.1784
	Neural Network	Neural Network	0.2181	0.2181

Best Model

Copyright © SAS Institute Inc. All rights reserved.

Deploying the Model

Copyright © SAS Institute Inc. All rights reserved.

Best Practices

Model Deployment

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

Copyright © SAS Institute Inc. All rights reserved.

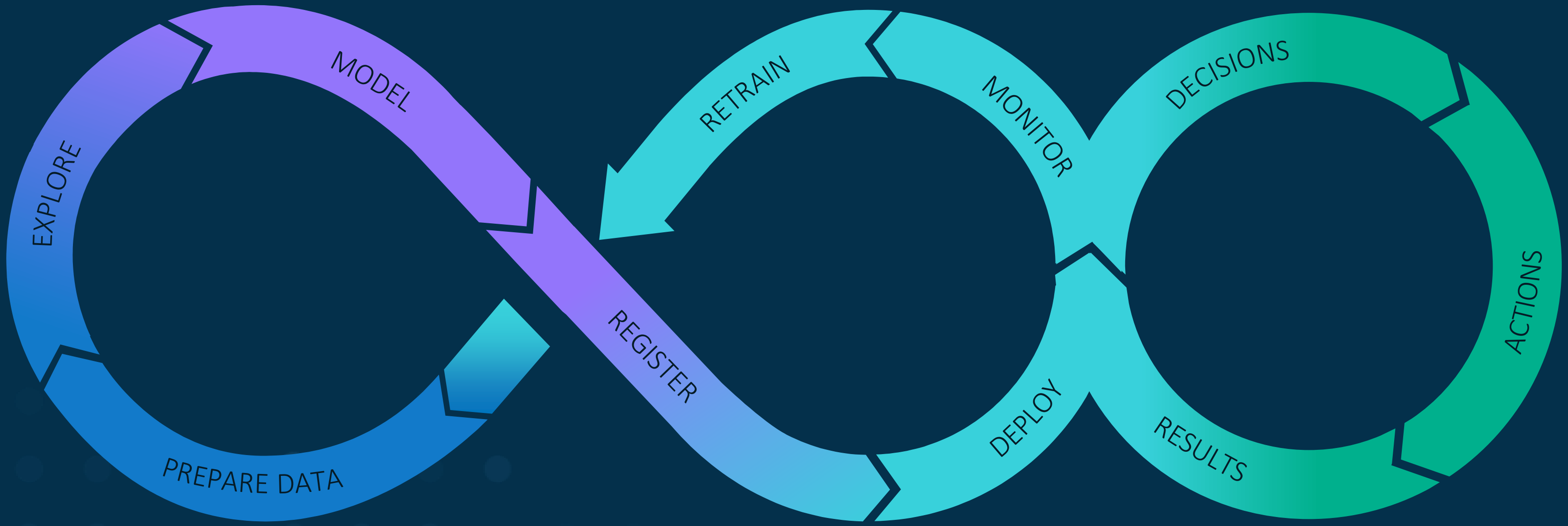
Best Practices

Model Deployment

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

Copyright © SAS Institute Inc. All rights reserved.

Analytics Operationalization with SAS



Copyright © SAS Institute Inc. All rights reserved.

Best Practices

Format of Presentation

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

Copyright © SAS Institute Inc. All rights reserved.

Best Practice

Be analytically savvy and creative

analytical creative

It's both science *and* art!

Copyright © SAS Institute Inc. All rights reserved.

Resources

Copyright © SAS Institute Inc. All rights reserved.

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

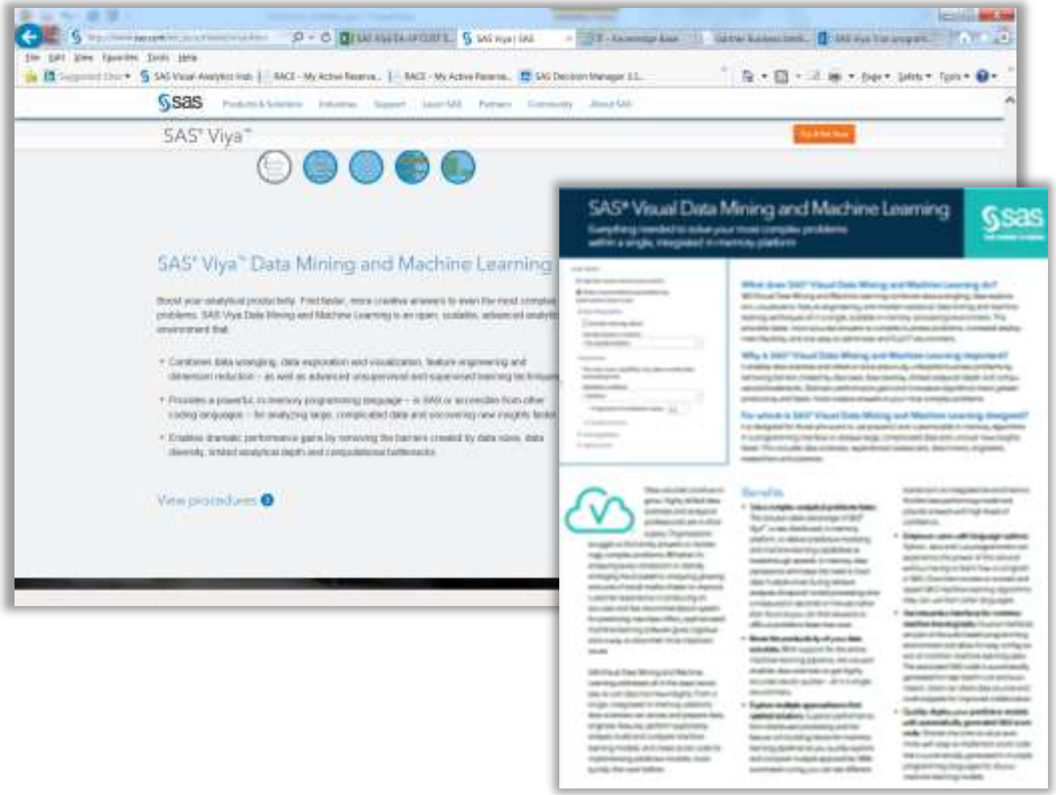
Copyright © SAS Institute Inc. All rights reserved.

SAS® Visual Data Mining and Machine Learning

Where to learn more?

Key Resources

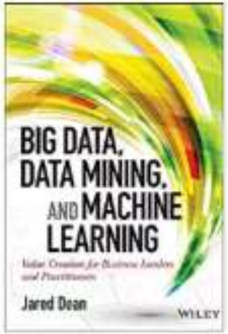
- SAS VDMML Product Web Page
- Factsheet
- SAS Viya Brochure
- Documentation
- VDMML SAS Community



Copyright © SAS Institute Inc. All rights reserved.

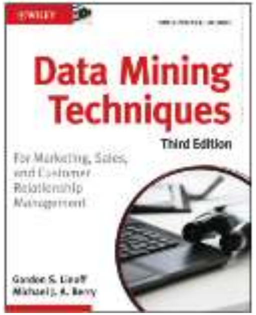
Resources

Suggested Reading



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

Available on [Amazon]

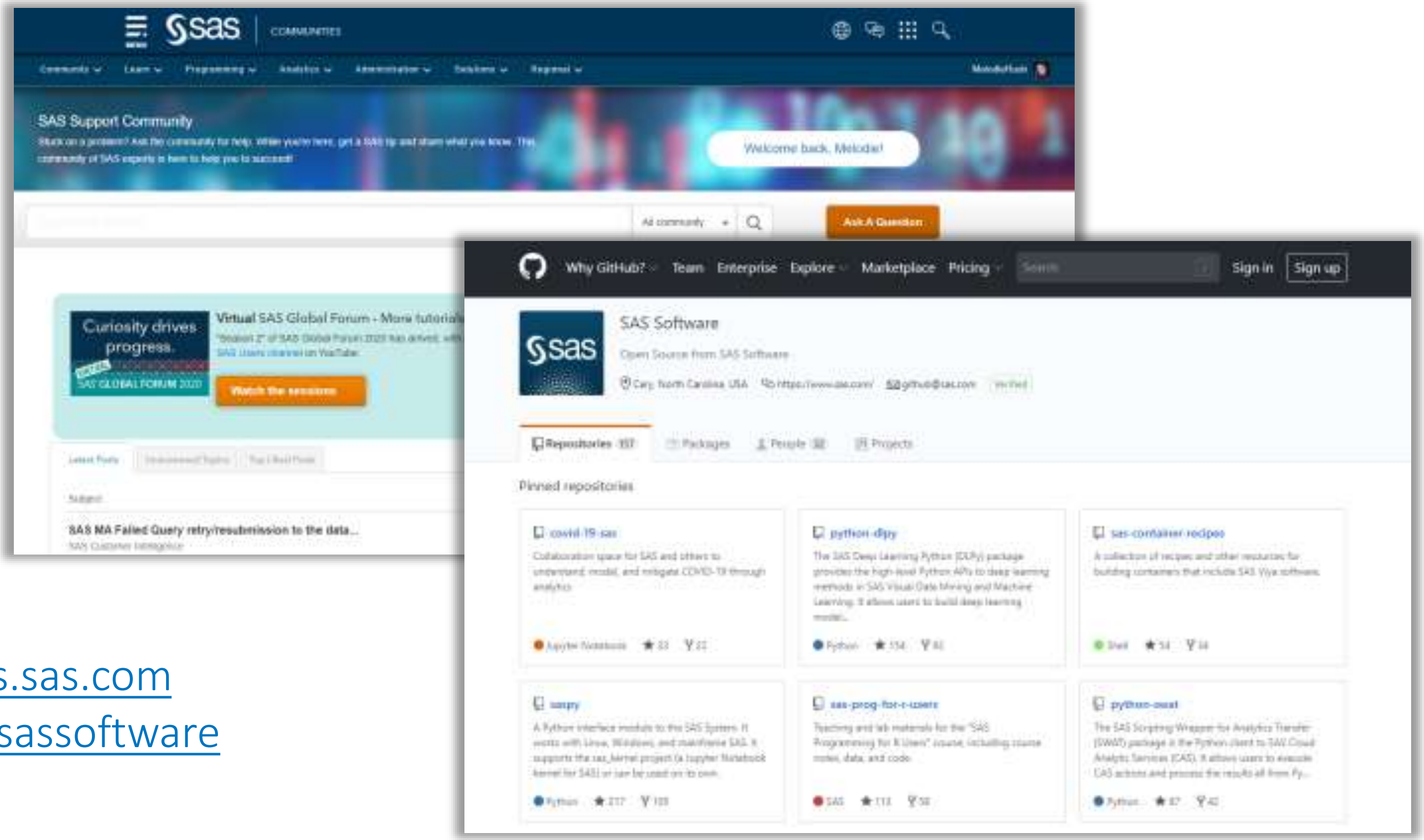


Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management
by Gordon S. Linoff and Michael J. A. Berry

Available on [Amazon]

Copyright © SAS Institute Inc. All rights reserved.

Communities



Communities.sas.com
Github.com/sassoftware

Copyright © SAS Institute Inc. All rights reserved.

Questions?

Thank you for your time and attention!

Connect with me:

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

Twitter: @Melodie_Rush

sas.com

Copyright © SAS Institute Inc. All rights reserved.