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

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

• 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

- Planning
- Project methodology
- Standards
- Continuous Process Improvement

TECHNOLOGY

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

BUSINESS PROCESS

- Vision & Leadership
- Team composition
- Enterprise authority

PEOPLE

DECISION ANALYTICS

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**
- Exploratory Analysis
- Descriptive Segmentation
- Predictive Modeling
- Model Validation & Registration

**DATA SCIENTIST**
- Model Validation
- Model Deployment
- Model Monitoring
- Data Preparation

**IT/SYSTEMS MANAGEMENT**
- Model Validation
- Model Deployment
- Model Monitoring
- Data Preparation

*Involve all the relevant people/roles*
Best Practice

Use the Technology and Method that 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.
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: \( Y = f(X_1, X_2, \ldots) \) \( \leftarrow \) fill-in the likely X’s
The SAS Platform
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
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
Essential Deployment Tasks

- Assess models
- Compare models
- Score the champion model
- Monitor model performance over time
- Update the model as needed
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
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.
Partitioning
Honest Assessment
## SAMPLE
### Data Partitioning

<table>
<thead>
<tr>
<th>PARTITION</th>
<th>ROLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>Used to fit the model</td>
</tr>
<tr>
<td>Validation</td>
<td>Used to validate the model and prevent over-fitting</td>
</tr>
<tr>
<td>Test</td>
<td>Used to provide unbiased estimate of model performance</td>
</tr>
</tbody>
</table>
WHAT IS OPTIMAL PARTITION?

- Training: 40%
- Validation: 30%
- Test: 30%
Best Practice
SAMPLE: Data Partitioning

WHAT IS OPTIMAL PARTITION?

- Training: 40%
- Validation: 0%
- Test: 60%
Best Practice
Sample: Data Partitioning

WHAT IS OPTIMAL PARTITION?

- Training: 70%
- Validation: 30%
- Test: 0%

It depends!
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?
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.
Selecting Target
Choosing your target

- Choosing the Target
- Response vs. Propensity
- Number of Models
Preparing Data
EXPLORE & MODIFY
Iterative Relationship with Data Preparation

Data Preparation

Data Modification

Data Exploration
Once you have an analytics-ready table:

- Examine *Categorical* Variables
- Examine *Continuous* Variables
- Explore *Missing* Values
- *Cluster* Variables
Explore & Modify

Categorical Variables

- Bar charts showing frequency of different categories of products: BED, CHAIR, DESK, SOFA, TABLE.
- Frequency distribution of taller visits.
- Distribution of CLUSTER_CODE showing frequency of various cluster codes.
- Frequency distribution of two-letter abbreviations for state names.
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

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
Variable Selection or Variable Reduction

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

Variable Selection

• Regression based
• Criterion Based
• Variable Screening
• Variable Clustering

Variable Combination

• Principal Components uncorrelated linear combinations of all input variables
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)
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**
## Summary of Data Preparation

<table>
<thead>
<tr>
<th>Topic</th>
<th>Common Challenges</th>
<th>Suggested Best Practices</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Collection</td>
<td>• Biased data&lt;br&gt;• Incomplete data&lt;br&gt;• High-dimensional data&lt;br&gt;• Sparsity</td>
<td>• Take time to understand the business problem and its context&lt;br&gt;• Enrich the data&lt;br&gt;• Dimension reduction (Feature Extraction, Variable Clustering, and Variable Selection)&lt;br&gt;• Change representation of data (Transformations node)</td>
</tr>
</tbody>
</table>
# Summary of Data Preparation

<table>
<thead>
<tr>
<th>Topic</th>
<th>Common Challenges</th>
<th>Suggested Best Practices</th>
</tr>
</thead>
</table>
| 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) |
## Summary of Data Preparation

<table>
<thead>
<tr>
<th>Topic</th>
<th>Common Challenges</th>
<th>Suggested Best Practices</th>
</tr>
</thead>
</table>
| 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)                |
# Summary of Data Preparation

<table>
<thead>
<tr>
<th>Topic</th>
<th>Common Challenges</th>
<th>Suggested Best Practices</th>
</tr>
</thead>
<tbody>
<tr>
<td>High-cardinality variables</td>
<td>• Overfitting&lt;br&gt;• Unknown categorical values in holdout data</td>
<td>• Binning (Transformations node)&lt;br&gt;• Replacement (Replacement node)</td>
</tr>
<tr>
<td>Missing Data</td>
<td>• Information loss&lt;br&gt;• Bias</td>
<td>• Binning (Transformations node)&lt;br&gt;• Imputation (Imputation node)</td>
</tr>
<tr>
<td>Strong multicollinearity</td>
<td>• Unstable parameter estimates</td>
<td>• Dimension reduction (Feature Extraction, Variable Clustering, and Variable Selection nodes)</td>
</tr>
</tbody>
</table>
Best Practices
Optimizing Data

Determining Data
Selecting Target
Preparing Variables
Developing & Delivering the Model
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?
Model Development

Try various techniques and combinations of techniques.
## Complete List of SAS Enterprise Miner Nodes

<table>
<thead>
<tr>
<th>SAMPLE</th>
<th>Append</th>
<th>Data Partition</th>
<th>File Import</th>
<th>Filter</th>
<th>Merge</th>
<th>Sample</th>
<th>Input Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXPLORE</td>
<td>Association</td>
<td>Cluster</td>
<td>Graph Explore</td>
<td>Variable Clustering</td>
<td>DMDB</td>
<td>Market Basket</td>
<td>Link Analysis</td>
</tr>
<tr>
<td>MODIFY</td>
<td>Drop</td>
<td>Impute</td>
<td>Interactive Binning</td>
<td>Principal Components</td>
<td>Replacement</td>
<td>Rules Builder</td>
<td>Transform Variables</td>
</tr>
<tr>
<td>MODEL</td>
<td>Decision Tree</td>
<td>AutoNeural Regression</td>
<td>Neural Network</td>
<td>Partial Least Squares</td>
<td>Dmne Regression</td>
<td>DM Neural Ensemble</td>
<td>Rule Induction</td>
</tr>
<tr>
<td></td>
<td>Incremental Response</td>
<td>Survival Analysis</td>
<td>Credit Scoring*</td>
<td>TS Correlation</td>
<td>TS Data Prep</td>
<td>TS Dimension Reduction</td>
<td>TS Decomp.</td>
</tr>
<tr>
<td>ASSESS</td>
<td>Cutoff</td>
<td>Decisions</td>
<td>Model Comparison</td>
<td>Score</td>
<td>Segment Profile</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UTILITY</td>
<td>Control Point</td>
<td>End Groups Start Groups</td>
<td>Open Source Integration</td>
<td>Reporter</td>
<td>Score Code Export</td>
<td>Metadata</td>
<td>SAS Code Ext Demo</td>
</tr>
</tbody>
</table>

*Requires Credit Scoring for SAS Enterprise Miner Add-on License
SAS Enterprise Miner
# SAS® Visual Data Mining and Machine Learning Capabilities

## Analytics
- Analytic Astore Scoring
- Data Step
- Data Transpose
- DS2
- Function Compiler
- FedSQL
- Frequency / Crosstab
- Imputation
- Model Assessment
- Sampling and Partitioning
- Sentiment Analysis
- Sequencing/Pathing Analysis
- Text Mining
- Variable Binning
- Variable Cardinality Analysis
- Variable Clustering
- Variable Selection
- Variable Summary

## Statistics
- Cox Proportional Hazards
- Decision Trees
- Design Matrix
- General Additive Models
- Generalized Linear Models
- Independent Component Analysis
- Clustering
  - *K-means and K-modes*
- Linear Regression
- Linear Mixed Models
- Logistic Regression
- Model-Based Clustering
- Model Scoring
- Nonlinear Regression
- Ordinary Least Squares Regression
- Partial Least Squares Regression
- Pearson Correlation
- Principal Component Analysis
  - *Kernel Principal Comp. Analysis*
- Quantile Regression
- Shewhart Control Chart Analysis

## Machine Learning
- Automated ML
- *Data Science Pilot*
- Model Composer
- Audio Data Processing
- Bayesian Networks
- Boolean Rules
- Cross Validation ML
- Deep Learning
  - *Convolutional NN*
  - *Deep Forward NN*
  - *Recurrent NN*
- Transfer Learning
- Factorization Machine
- Frequent Item Set Mining
- Gaussian Mixture Model
- Gaussian Process Regression
- Gradient Boosting
- Hyperparameter Auto-tuning
- K Nearest Neighbor
- Image Processing *(incl. Biomedical)*
- Market Basket Analysis
- Model Interpretability
  - LIME, ICE, PD, Shapley
- Moving Windows PCA
- Multitask Learning
- Network Analytics
- Neural Networks
- Random and Isolation Forests
- Recommendation Engine
- Robust PCA
- Semi-supervised Learning
- Sparse Machine Learning
- Support Vector Data Description
- Support Vector Machine
- *t*-distributed SNE
- Text Parsing

*Last updated 12/10/2019*
*For documentation please click [here](#)*
SAS Visual Data Mining and Machine Learning
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
Choosing a Model
Model Selection

• Evaluate model metrics
• Consider business knowledge
• Recognize constraints
How do we choose?

<table>
<thead>
<tr>
<th>Prediction Type</th>
<th>Validation Fit Statistic</th>
<th>Direction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decisions</td>
<td>Misclassification</td>
<td>smallest</td>
</tr>
<tr>
<td></td>
<td>Average Profit/Loss</td>
<td>largest/smallest</td>
</tr>
<tr>
<td></td>
<td>Kolmogorov-Smirnov Statistic</td>
<td>largest</td>
</tr>
<tr>
<td>Rankings</td>
<td>ROC Index (concordance)</td>
<td>largest</td>
</tr>
<tr>
<td></td>
<td>Gini Coefficient</td>
<td>largest</td>
</tr>
<tr>
<td>Estimates</td>
<td>Average Squared Error</td>
<td>smallest</td>
</tr>
<tr>
<td></td>
<td>Schwarz's Bayesian Criterion</td>
<td>smallest</td>
</tr>
<tr>
<td></td>
<td>Log-Likelihood</td>
<td>largest</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th>AIC</th>
<th>Captured Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASE</td>
<td>KS Statistic</td>
</tr>
<tr>
<td>MSE</td>
<td>Misclassification</td>
</tr>
<tr>
<td>ROC</td>
<td>Average Profit/Loss</td>
</tr>
<tr>
<td>Gain</td>
<td>Cumulative Lift</td>
</tr>
<tr>
<td>Lift</td>
<td>Cumulative Captured Response</td>
</tr>
<tr>
<td>Gini</td>
<td>Cumulative Percent Response</td>
</tr>
</tbody>
</table>

Available for training, validation and test datasets
Assess Cumulative charts
Assess Non-Cumulative charts
SAS® Enterprise Miner™
Model Comparison Node

<table>
<thead>
<tr>
<th>Selected Model</th>
<th>Predecessor Node</th>
<th>Model Node</th>
<th>Model Description</th>
<th>Target Variable</th>
<th>Target Label</th>
<th>Selection Criterion: Valid: Misclassification Rate ▲</th>
<th>Train: Misclassification Rate</th>
<th>Valid: Lift</th>
<th>Train: Schwarz’s Bayesian Criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>Reg4</td>
<td>Reg4</td>
<td>Regression DT</td>
<td>TARGET…</td>
<td>Donated…</td>
<td>0.249441</td>
<td>0.24965</td>
<td>1.539784</td>
<td>15059.33</td>
</tr>
<tr>
<td>HPDMForest</td>
<td>HPDMForest</td>
<td>HP Forest</td>
<td>Regression stepwise</td>
<td>TARGET…</td>
<td>Donated…</td>
<td>0.248957</td>
<td>0.247979</td>
<td>1.429799</td>
<td></td>
</tr>
<tr>
<td>HPReg4</td>
<td>HPReg4</td>
<td>HP Regression stepwise</td>
<td>TARGET…</td>
<td>Donated…</td>
<td>0.250473</td>
<td>0.249428</td>
<td>1.546658</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reg5</td>
<td>Reg5</td>
<td>Regression PC</td>
<td>Regression PC</td>
<td>TARGET…</td>
<td>Donated…</td>
<td>0.250645</td>
<td>0.249133</td>
<td>1.443547</td>
<td></td>
</tr>
<tr>
<td>HPReg3</td>
<td>HPReg</td>
<td>HP Reg - Backward</td>
<td>Regression PC</td>
<td>TARGET…</td>
<td>Donated…</td>
<td>0.250817</td>
<td>0.249281</td>
<td>1.457295</td>
<td></td>
</tr>
<tr>
<td>Reg2</td>
<td>Reg2</td>
<td>Regression Forward</td>
<td>Regression Forward</td>
<td>TARGET…</td>
<td>Donated…</td>
<td>0.251161</td>
<td>0.247585</td>
<td>1.361059</td>
<td>15075.82</td>
</tr>
<tr>
<td>Reg3</td>
<td>Reg3</td>
<td>Regression Stepwise</td>
<td>Regression Stepwise</td>
<td>TARGET…</td>
<td>Donated…</td>
<td>0.251161</td>
<td>0.247585</td>
<td>1.361059</td>
<td>15075.82</td>
</tr>
<tr>
<td>Reg</td>
<td>Reg</td>
<td>Regression Backward</td>
<td>Regression Backward</td>
<td>TARGET…</td>
<td>Donated…</td>
<td>0.251849</td>
<td>0.247732</td>
<td>1.361059</td>
<td>15075.56</td>
</tr>
<tr>
<td>HPReg2</td>
<td>HPReg2</td>
<td>HP Reg Fast Backward</td>
<td>Regression 2 Poly</td>
<td>TARGET…</td>
<td>Donated…</td>
<td>0.252193</td>
<td>0.248838</td>
<td>1.539784</td>
<td></td>
</tr>
<tr>
<td>Reg8</td>
<td>Reg8</td>
<td>Regression 2 Poly</td>
<td>Regression Full</td>
<td>TARGET…</td>
<td>Donated…</td>
<td>0.253226</td>
<td>0.246478</td>
<td>1.484792</td>
<td>15017.38</td>
</tr>
<tr>
<td>Reg6</td>
<td>Reg6</td>
<td>Regression Full</td>
<td>Regression Full</td>
<td>TARGET…</td>
<td>Donated…</td>
<td>0.253398</td>
<td>0.246773</td>
<td>1.622272</td>
<td>15639.84</td>
</tr>
<tr>
<td>Reg9</td>
<td>Reg9</td>
<td>Reg 2-way Int 2 Poly</td>
<td>Regression 2-way Interactions</td>
<td>TARGET…</td>
<td>Donated…</td>
<td>0.258214</td>
<td>0.241463</td>
<td>1.429799</td>
<td>16427.17</td>
</tr>
<tr>
<td>Reg7</td>
<td>Reg7</td>
<td>Regression 2-way Interactions</td>
<td>Regression 2-way Interactions</td>
<td>TARGET…</td>
<td>Donated…</td>
<td>0.295544</td>
<td><strong>0.21211</strong></td>
<td>1.127342</td>
<td>33523.52</td>
</tr>
</tbody>
</table>

SAS Enterprise Miner assumes decision processing and selects the model with the lowest **misclassification rate** when there is a binary target.
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
Assess
Cumulative charts
Assess
Non-cumulative charts

Lift Reports
Response Percentage

Response Percentage

Depth

ROC Reports
Accuracy

Accuracy

Cutoff
### SAS® Visual Data Mining and Machine Learning

**Model Comparison Node**

#### Model Comparison Results

<table>
<thead>
<tr>
<th>Champion</th>
<th>Name</th>
<th>Algorithm Name</th>
<th>Misclassification Rate (Event)</th>
<th>Misclassification Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Best Model" /></td>
<td>GB Tune Explain</td>
<td>Gradient Boosting</td>
<td><strong>0.0699</strong></td>
<td><strong>0.0699</strong></td>
</tr>
<tr>
<td></td>
<td>Gradient Boosting</td>
<td>Gradient Boosting</td>
<td>0.0912</td>
<td>0.0912</td>
</tr>
<tr>
<td></td>
<td>Ensemble</td>
<td>Ensemble</td>
<td>0.1549</td>
<td>0.1549</td>
</tr>
<tr>
<td></td>
<td>Forward Logistic Regression</td>
<td>Logistic Regression</td>
<td>0.1711</td>
<td>0.1711</td>
</tr>
<tr>
<td></td>
<td>Stepwise Logistic Regression</td>
<td>Logistic Regression</td>
<td>0.1711</td>
<td>0.1711</td>
</tr>
<tr>
<td></td>
<td>Decision Tree</td>
<td>Decision Tree</td>
<td>0.1784</td>
<td>0.1784</td>
</tr>
<tr>
<td></td>
<td>Neural Network</td>
<td>Neural Network</td>
<td>0.2181</td>
<td>0.2181</td>
</tr>
</tbody>
</table>

**Best Model**

- GB Tune Explain

---

Copyright © SAS Institute Inc. All rights reserved.
Deploying the Model
Best Practices

Model Deployment

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

• Incorporate and share knowledge
• Automate ETL (Extract, Transform, Load)
• Automate process
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](http://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
Key Resources

- [SAS VDMML Product Web Page](#)
- [Factsheet](#)
- [SAS Viya Brochure](#)
- [Documentation](#)
- [VDMML SAS Community](#)
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
Communities

Communities.sas.com
Github.com/sassoftware
Questions?

Thank you for your time and attention!

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