Demo Decision Tree (Autonomous)

Dr. Goutam Chakraborty

SAS EM Demo Procedure

- Assess the Maximal Tree using validation data
  - Select the **Decision Tree** node. In the Decision Tree node's Train properties, change the **Use Frozen Tree** property's value from No to **Yes**.
  - Right-click the **Decision Tree** node and run it. Select **Results**.
  - Select **View** ⇨ **Model** ⇨ **Subtree Assessment Plot**.
  - To further explore validation performance, select the arrow in the upper left corner of the Subtree Assessment Plot, and switch the assessment statistic for example to **Misclassification Rate**
  - Right-click the **Decision Tree** node and select **Rename**. Name the node **Maximal Tree**

SAS EM Demo Procedure (Continued)

Building a Tree autonomously

- Drag another **Decision Tree** node from the Model tab Right-click the **Decision Tree** node and run it. Select **Results**.
- Rename the Tree as Optimal (Misclassification) Tree.

Property	Value
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Decision
Assessment Fraction	0.25
Cross Validation	
Perform Cross Validation	No
Number of Subsets	10
Number of Repeats	1
Seed	12345

Metadata plays an important role in how SAS Enterprise Miner functions. Recall that a binary target variable was selected for the project. Based on this, SAS Enterprise Miner assumes that you want a tree that is optimized for making the best ***decisions*** (as opposed to the best rankings or best probability estimates). That is, under the current project settings, SAS Enterprise Miner chooses, by default, the tree with the lowest misclassification rate on the validation sample

What if I Want to Use ASE (Average Square Error)?

$$Average\ square\ error = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

This is easy to understand for interval targets

- What do we do with binary or multiple category targets?

$$Average\ square\ error = \frac{1}{N \cdot L}\sum_{i=1}^{N}\sum_{j=1}^{L}\left(I\left(y_i = C_j\right) - \hat{p}_{ij}\right)^2$$

$$I\left(y_i = C_j\right) = \begin{Bmatrix} 1 & y_i = C_j \\ 0 & y_i \neq C_j \end{Bmatrix}$$

SAS EM Demo Procedure (Continued)

- What if our prediction objective is not decision but estimate of probability for donation?
  - Navigate to the **Model** tab. Drag a new **Decision Tree** and name it as Probability Tree
  - In the properties panel, change Assessment Measure to **Average Square Error**
  - Run this Tree and compare results of the Probability Tree with the Misclassification Tree.
- Explore different panels of the results
  - Treemap
  - Tree
  - Leaf statistic bar chart
  - Variable importance
  - Score rankings overlay
  - Fit statistics
  - Output window

More on Decision Tree Results

- Select **View ⇨ Model ⇨ Variable Importance**

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
GiftCnt36	Gift Count 3...	1	1.0000	1.0000	1.0000
GiftAvgLast	Gift Amount...	1	0.5241	0.6719	1.2819
DemMedH...	Median Ho...	1	0.5028	0.1086	0.2160
GiftTimeLast	Time Since ...	1	0.4809	0.4453	0.9259
DemPctVet...	Percent Vet...	0	0.0000	0.0000	.
GiftAvg36	Gift Amount...	0	0.0000	0.0000	.
GiftAvgCard...	Gift Amount...	0	0.0000	0.0000	.
DemAge	Age	0	0.0000	0.0000	.
GiftAvgAll	Gift Amount...	0	0.0000	0.0000	.
GiftCntAll	Gift Count A...	0	0.0000	0.0000	.
GiftCntCard...	Gift Count ...	0	0.0000	0.0000	.
GiftCntCard...	Gift Count ...	0	0.0000	0.0000	.
GiftTimeFirst	Time Since ...	0	0.0000	0.0000	.
PromCntCa...	Promotion ...	0	0.0000	0.0000	.
PromCnt12	Promotion ...	0	0.0000	0.0000	.
PromCnt36	Promotion ...	0	0.0000	0.0000	.
PromCntAll	Promotion ...	0	0.0000	0.0000	.
PromCntCa...	Promotion ...	0	0.0000	0.0000	.

Self Study (See Handout)

- Try changing **number of branches** (from default of 2 to say 3) and explore what effect if any that has on your results
  - Try changing exhaustive search size limit along with maximum branches and explore what effect if any that has on your results
- Try changing **how splits are evaluated** (default is Chi-square to Gini and Entropy) and explore what effect if any that has on your results
- Try **different pruning options** (combinations of Subtree method, Assessment measure) and explore what effect if any that has on your results
- Try building **different sizes of the tree** (combinations of logworth threshold, maximum tree depth, minimum leaf size and threshold depth adjustment) and explore what effect if any that has on your results

My thoughts about learning tool options

- For predictive modeling certification exam, you do need to be fairly familiar with the tools/options/properties in SAS EM
- For your individual research project or other projects, you should try to play with all the options systematically and explore what effects that produce on your results
- For building your knowledge:
  - Understand concepts and theories that have been encoded in the tool (such as log-worth, variable importance, etc.)
  - Take the same data and try to replicate SAS EM results using Python library of your choice (or, R) and understand differences and similarities