Data Mining and Business Intelligence

Lecture 3: SAS Enterprise Miner Intro

Jing Peng
University of Connecticut

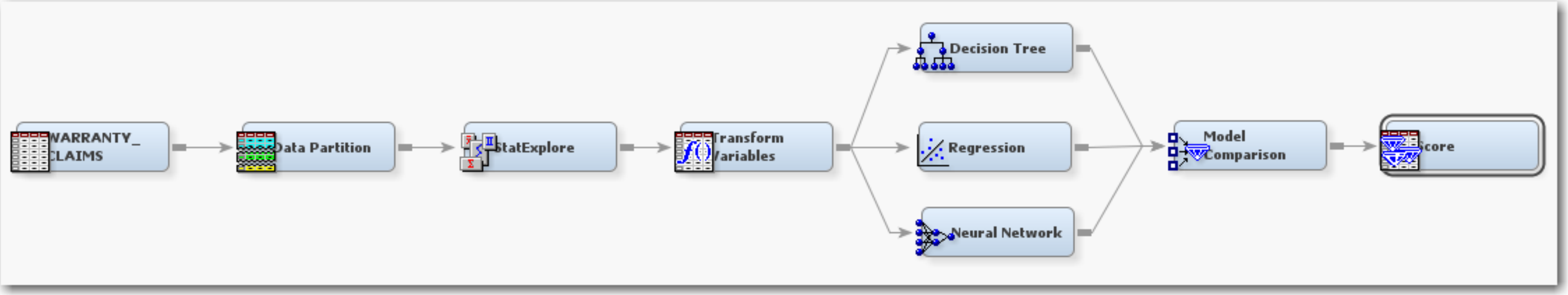
2/5/20

Recap via Group Discussions

- In high dimensional spaces, what happens to Euclidean distance and cosine similarity
  - What are the intuitions behind?
  - What are the consequences?
- Filter, wrapper, embedded feature selection models
  - Definitions
  - Pros and cons
- Intuitively, why Lasso tends to give more sparse parameter estimates than Ridge?

SAS Enterprise Miner

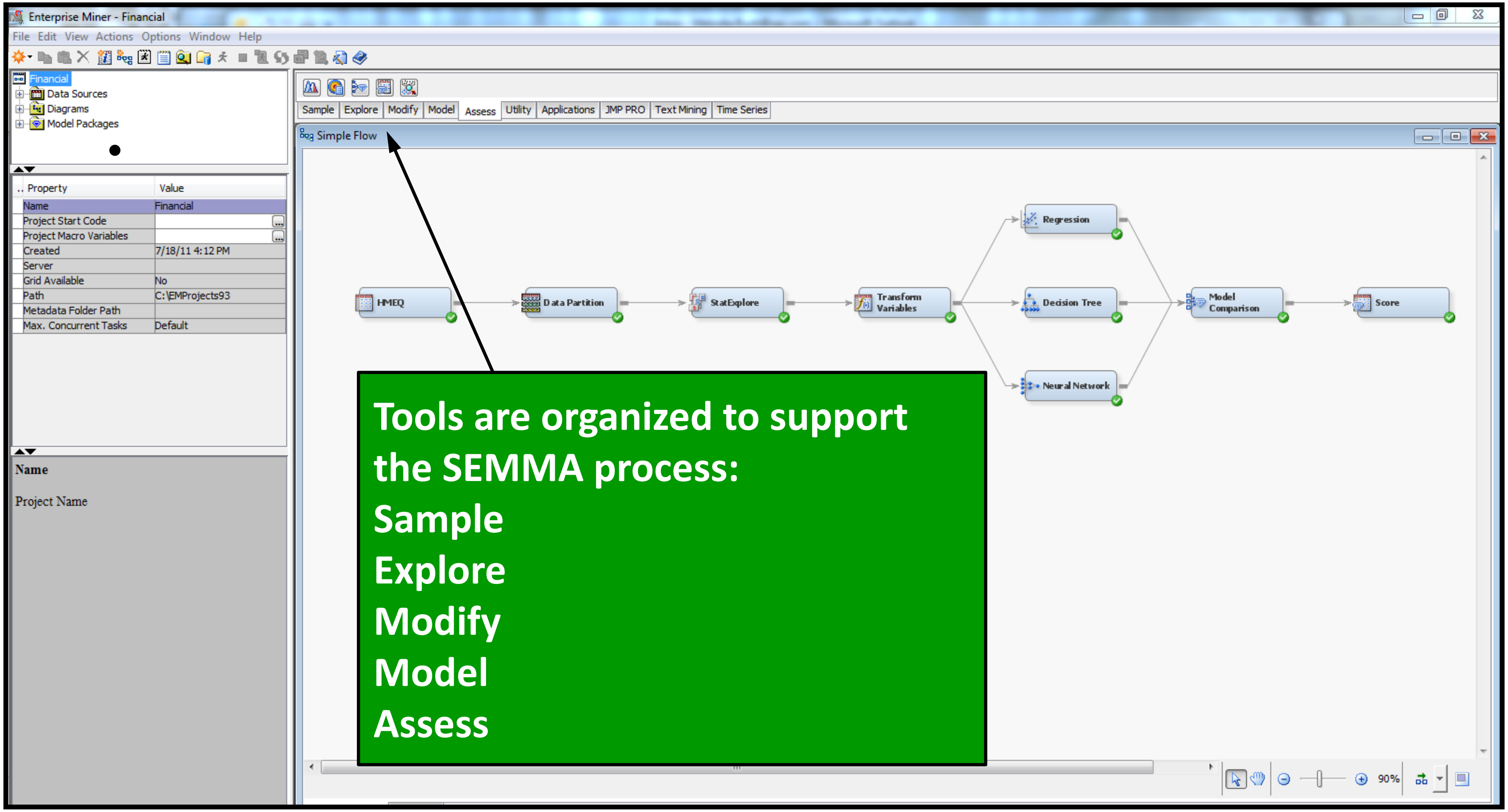
Data Mining Process in SAS



SAMPLE EXPLORE MODIFY MODEL ASSESS and SCORE

Sample Explore Modify Model Assess

Data Mining Process in SAS



Property	Value
Name	Financial
Project Start Code	
Project Macro Variables	
Created	7/18/11 4:12 PM
Server	
Grid Available	No
Path	C:\EMProjects93
Metadata Folder Path	
Max. Concurrent Tasks	Default

Name

Project Name

Tools are organized to support the SEMMA process:

Sample

Explore

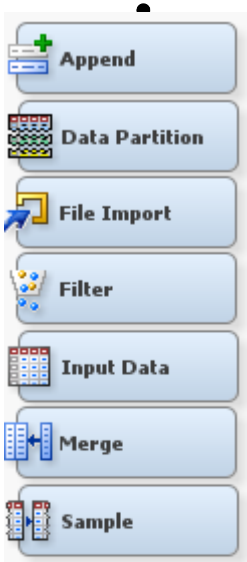
Modify

Model

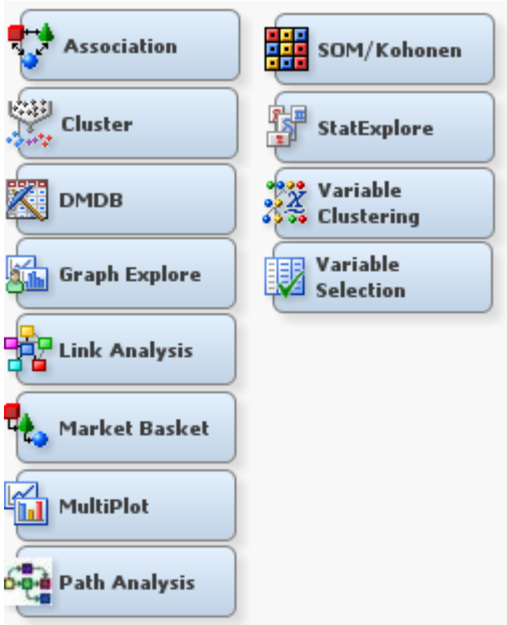
Assess

Data Mining Process in SAS

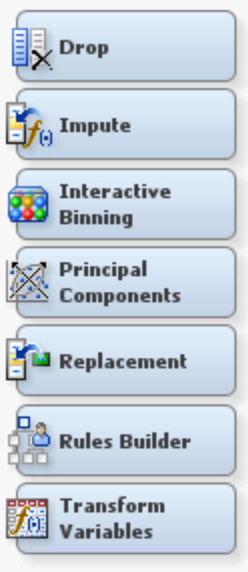
Sample



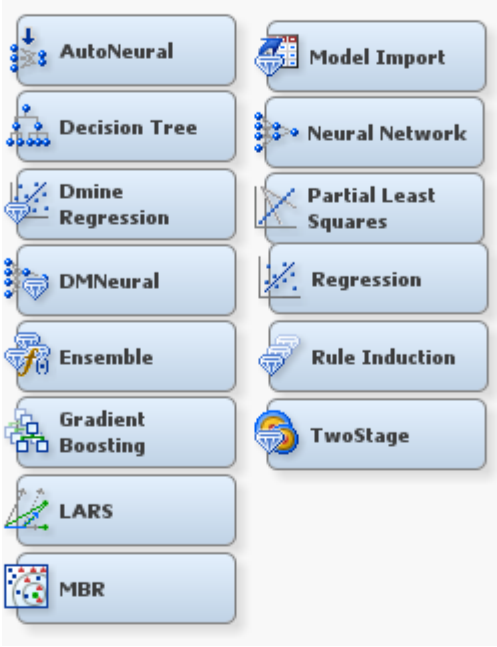
Explore



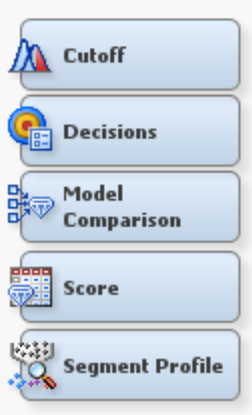
Modify



Model



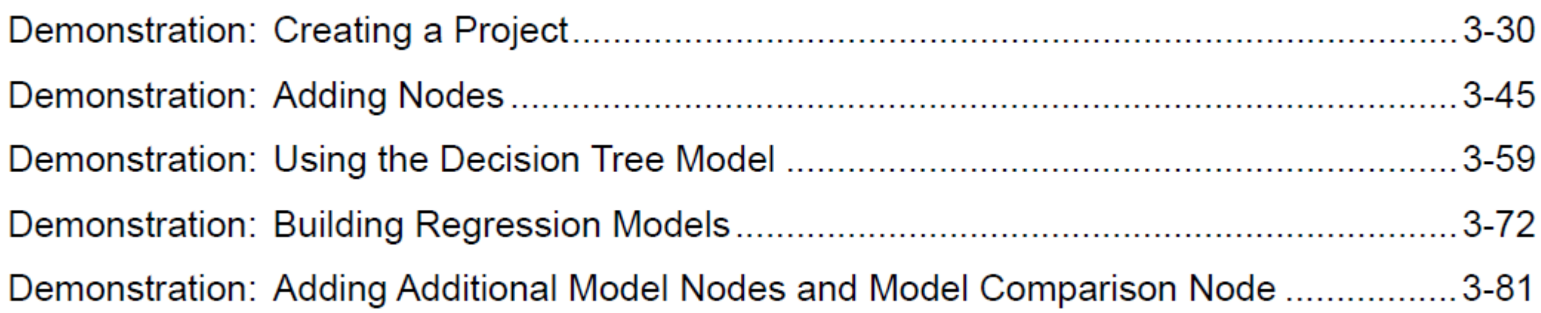
Assess



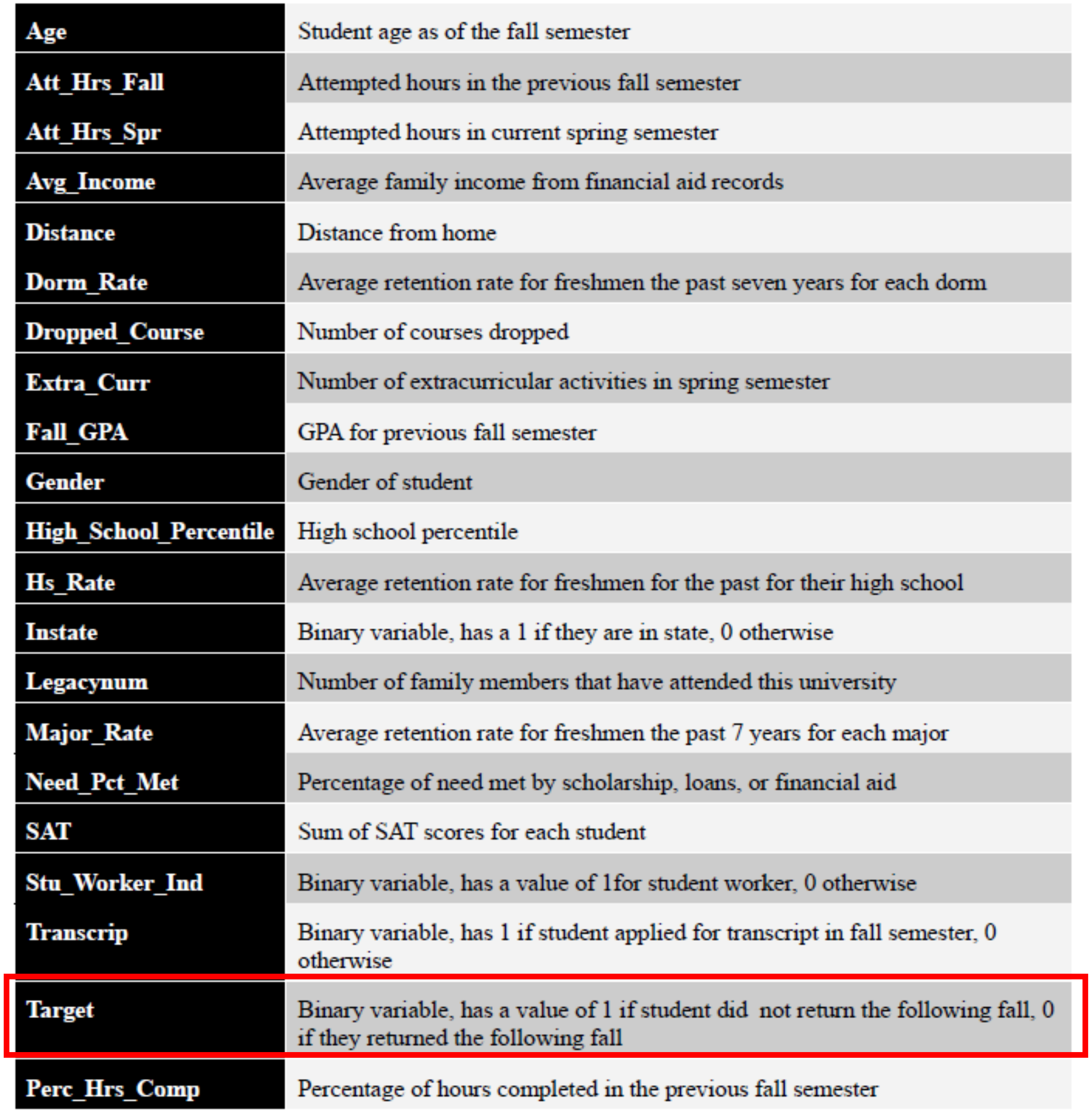
Retention Case Study

Develop a retention classification model to identify students who are likely to leave

ABA Chapter 3.4



Data Description

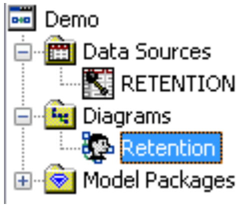


Age	Student age as of the fall semester
Att_Hrs_Fall	Attempted hours in the previous fall semester
Att_Hrs_Spr	Attempted hours in current spring semester
Avg_Income	Average family income from financial aid records
Distance	Distance from home
Dorm_Rate	Average retention rate for freshmen the past seven years for each dorm
Dropped_Course	Number of courses dropped
Extra_Curr	Number of extracurricular activities in spring semester
Fall_GPA	GPA for previous fall semester
Gender	Gender of student
High_School_Percentile	High school percentile
Hs_Rate	Average retention rate for freshmen for the past for their high school
Instate	Binary variable, has a 1 if they are in state, 0 otherwise
Legacynum	Number of family members that have attended this university
Major_Rate	Average retention rate for freshmen the past 7 years for each major
Need_Pct_Met	Percentage of need met by scholarship, loans, or financial aid
SAT	Sum of SAT scores for each student
Stu_Worker_Ind	Binary variable, has a value of 1for student worker, 0 otherwise
Transcrip	Binary variable, has 1 if student applied for transcript in fall semester, 0 otherwise
Target	Binary variable, has a value of 1 if student did not return the following fall, 0 if they returned the following fall
Perc_Hrs_Comp	Percentage of hours completed in the previous fall semester

Tips

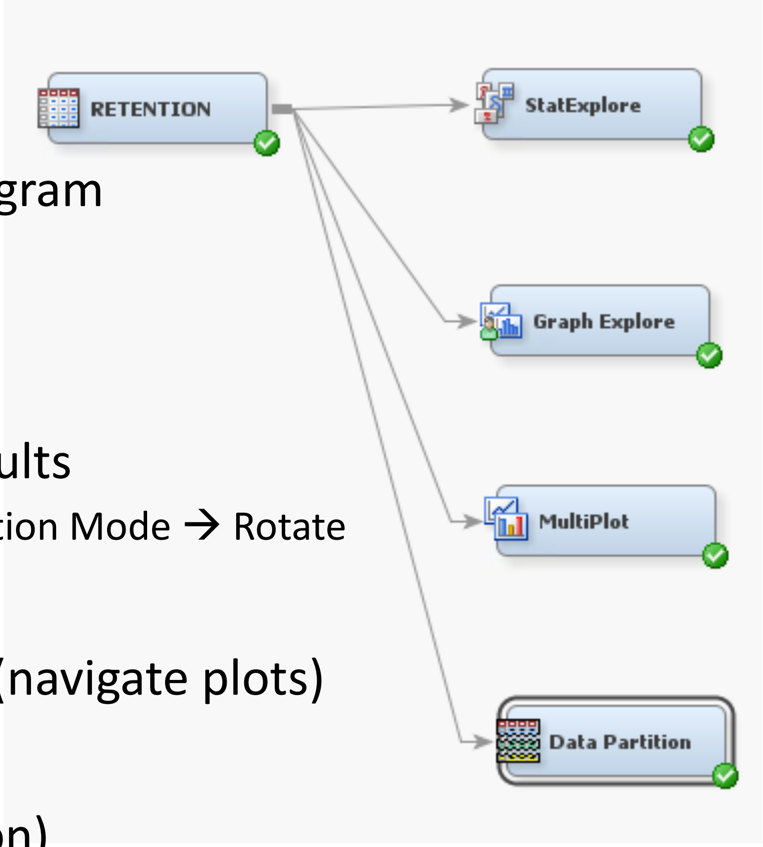
- SAS Enterprise Miner generates a huge volume of meta data (several GBs or more)
- Since P drive is a cloud drive, the data transfer between P drive and the virtual machine can become a painful bottleneck
- Recommendation: set the project location to the C drive of the virtual machine and copy it to P drive or shared folder once you are done.
- Caution: The files in C drive will be wiped out once you log off the VM.

Creating A Project


Demo
Data Sources
RETENTION
Diagrams
Retention
Model Packages

- New Project & New Diagram & New Library
- Create Data Source (“Advanced → Customized → Class Threshold 2” in step 4)

Adding Nodes

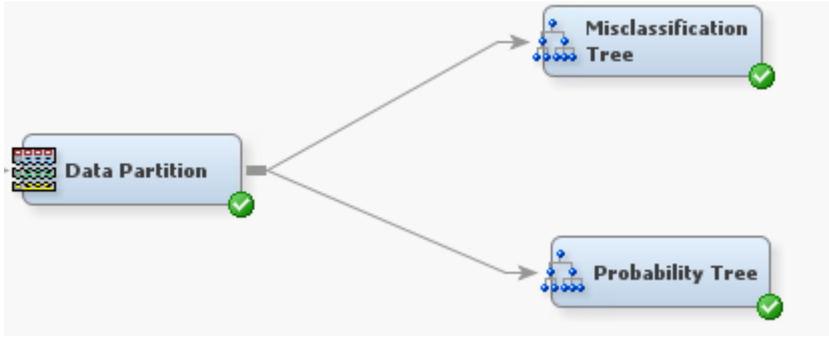


- Drag **Retention** and **StatExplore** nodes to diagram
  - Link them
  - Run StatExplore and examine results
- **GraphExplore**: connect, run and examine results
  - 3D Charts: X(Fall_GPA), Y(avg_income), Z(SAT) → Action Mode → Rotate
- **Multiplot**: connect, run and examine results (navigate plots)
- **Data Partition** Node: (60 train vs. 40 validation)

Decision Tree Model

- Misclassification Tree (Assessment Measure: Misclassification rate)
  - Rename, connect, and run
  - Tree Panel
  - View → Model → Subtree Assessment Plot
  - False positive and false negative

- Probability Tree (Assessment Measure: Average Squared Error)
  - Rename, change assessment measure, connect, and run
  - Compare: tree panel



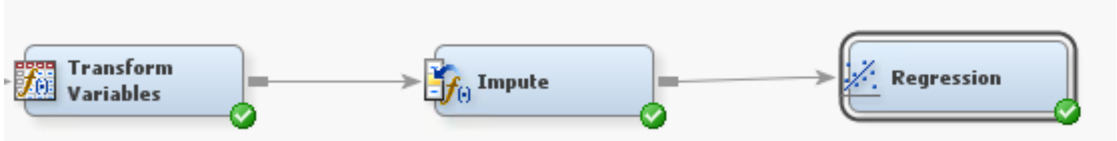
Regression Model

- **Retention** → Variables → All Interval variables → Explore

- Append **Transform** node to **Data Partition** node
  - Variables → All Interval variables→ Method → Max. Normal
  - Run → Skewness (original vs. computed)
  - Exported Data → Train → Explore → Plot → Histogram of SQRT_Distance

- Append **Impute** node to **Transform** node
  - Class/Interval Variables Input Method → Tree Surrogate
  - Indicator Variables (Type: Unique; Role: Input)

- Append **Regression** node to **Impute** node
  - Model Selection (selection model: stepwise; selection criterion: validation error; use selection defaults: no; ellipsis: entry-1.0, stay-0.5, max steps-30)
  - Run and examine results
  - View → Model → Iteration Plot
  - Output (bottom): message, parameter estimates, and classification table



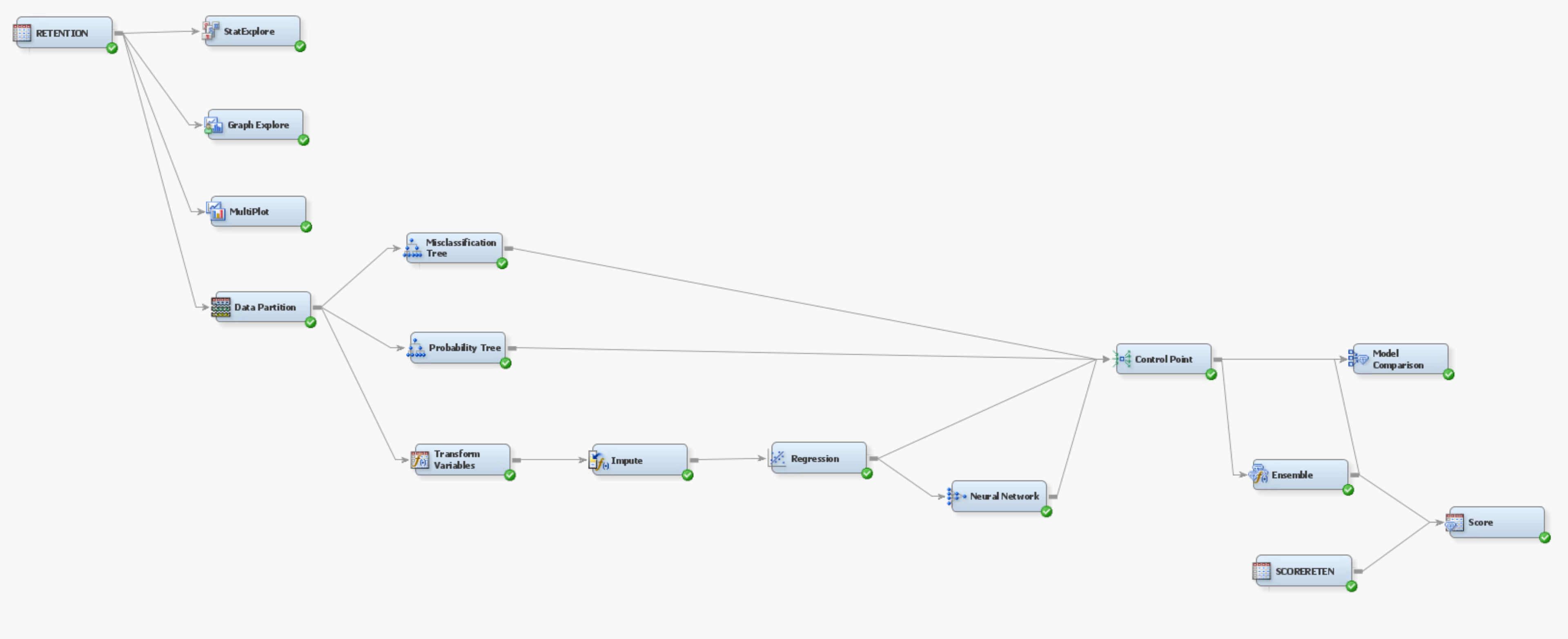
Model Comparison Node

- Append **Neural Network** node to Regression node
  - Train → Optimization → Preliminary Training → Enabled → No
  - Model Selection Criterion → Average Error
  - Run and examine results

- Add **Control Point** (Utility), **Ensemble** (Model), and **Model Comparison** (Assess) nodes
  - Links: all models to control point, control point to ensemble, control point and ensemble to model comparison
  - Model selection → Selection Statistic → Average Squared Error, Selection Table → Validation

- Use **Ensemble** node to **score** new data
  - Create Data Source for score dataset (step 7, role: score)
  - Drag score data to diagram and add **Score** node
  - Exported Data → Score → Explore → Plot → Bar → EM_Classification (prediction for target, role: category)
  - To get identical results with textbook: Options→ Preferences → Interactive Sampling → Fetch Size → Max

Final Diagram



References

- ABA Chapter 3.4
- **SAS Enterprise Miner Official Tutorial Videos**
- SAS Enterprise Miner Documentation: (press F1)