Data Mining in Health Analytics and A Quick Look at SAS Enterprise Miner Interface

Ram Poudel

Department of Behavioral Pediatrics

June 2nd , 2015

Objectives

- Overview Data Mining
- Overview the basic principle and best practices in Data Mining
- Describe the basic navigation of SAS EM
- Give a high level overview of three widely used modeling algorithms
- Discuss the application of Data Mining in health care.

Health Analytics and Categories

- Descriptive → Describing what has happened
- Predictive → Predicting what will happen
- Prescriptive → Determining what to do about

Data Mining

Definition

Data Mining is the analysis of large data sets to discover patterns and use those patterns to forecast or predict the likelihood of future events [1].

Patterns should be

Valid

Novel

Useful

Understandable

[1] Crockett, et. al, 2014

Modeling Essentials

Predict New Cases

Select useful inputs

Optimize complexity

Types of Prediction

Decisions

Rankings

Estimates

Honest Assessment: A Basic Principle of Data Mining

- Splitting the data:

Training Data Set – this is a must do
Validation Data Set – this is a must do
Testing Data Set – this is optional

Best Practices in Data Mining

- Handling Missing Values

  Empty vs. Missing

  1. Decision Trees have built in methods for handling missing values.

  2. Equation "type" algorithms, e.g. Logistic Regression and Neural Networks, do Complete Case Analysis

Best Practices in Data Mining

- Transformation of the variables

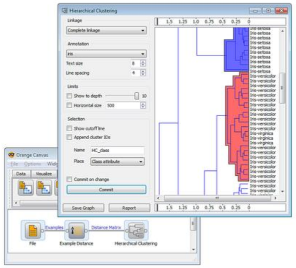
  For Regression, it is a must but better for Decision Tree and Neural Net too.

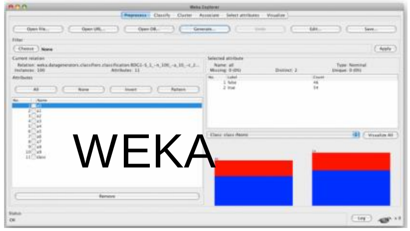
- Variable Selection

  Better to select variables especially for Neural Net

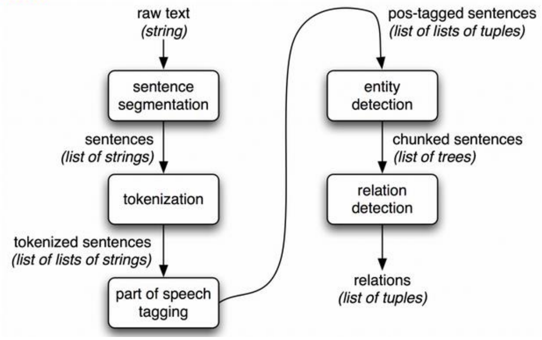
Open Tools for Data Mining

Orange

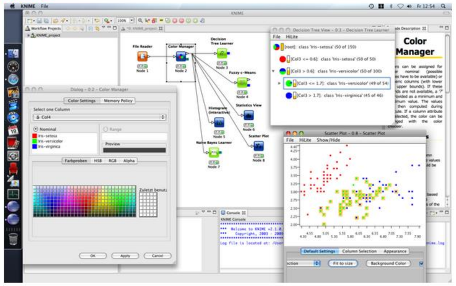




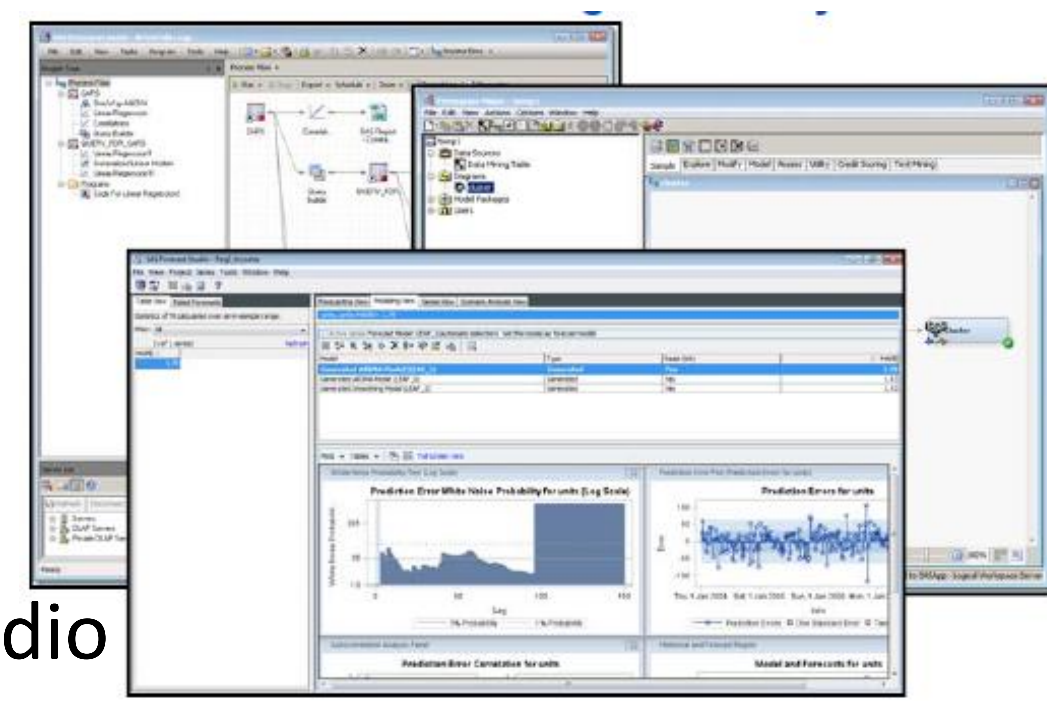
NLTK



KNIME



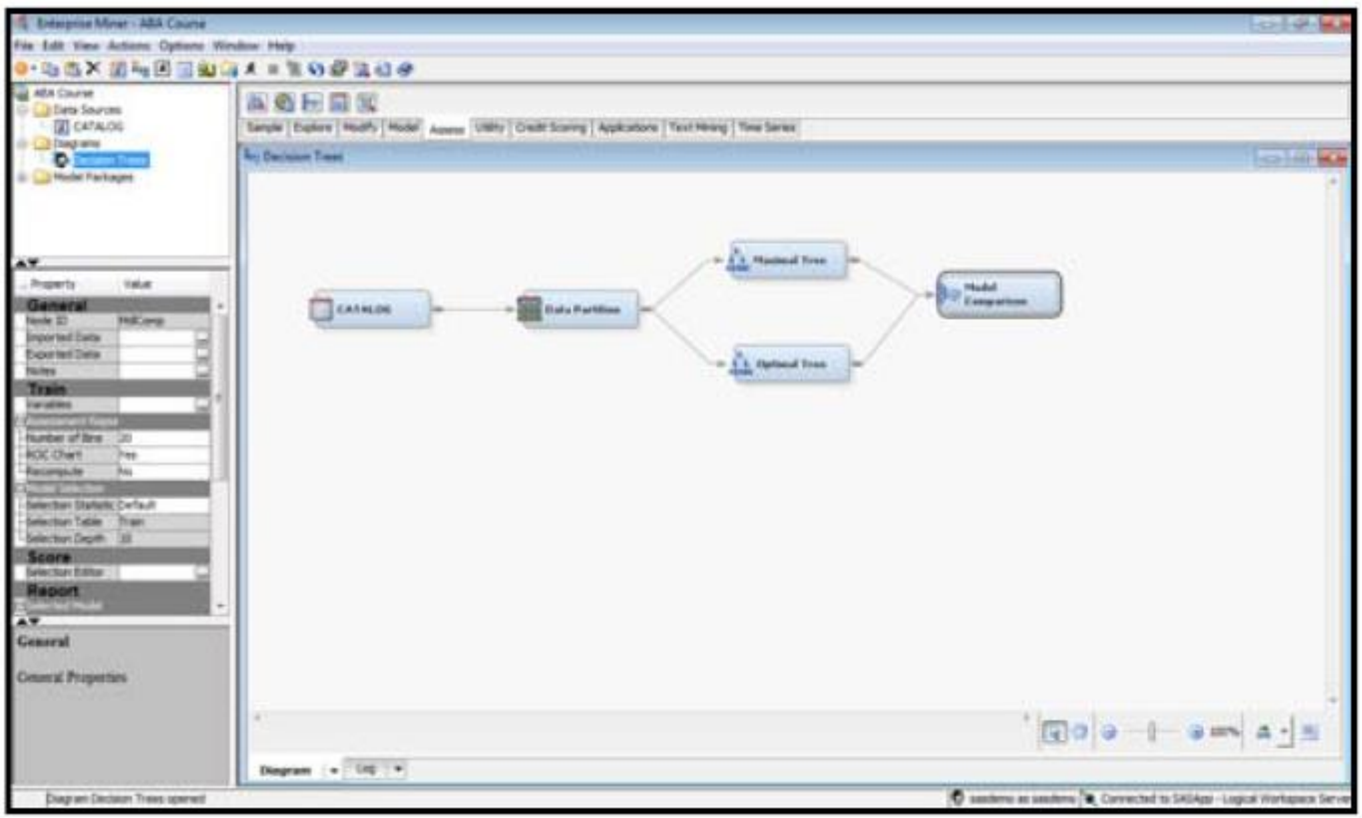
Tools for Data Mining from SAS System



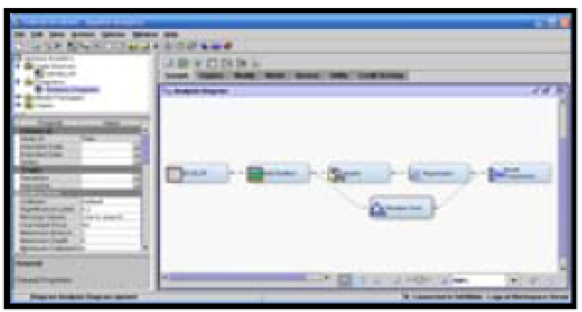
- SAS EG
- SAS EM
- SAS Forecast Studio
- JMP

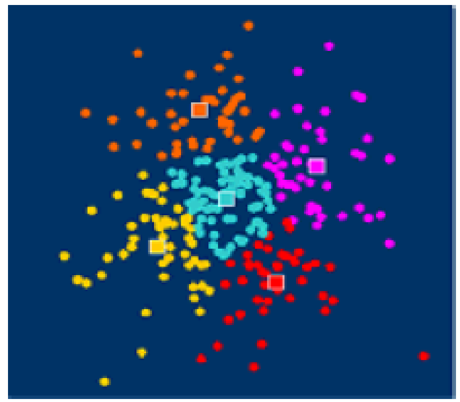
Introduction to SAS Enterprise Miner

SAS EM streamlines the Data Mining process to create highly accurate predictive and descriptive models based on the vast amount of data gathered from across an entity.

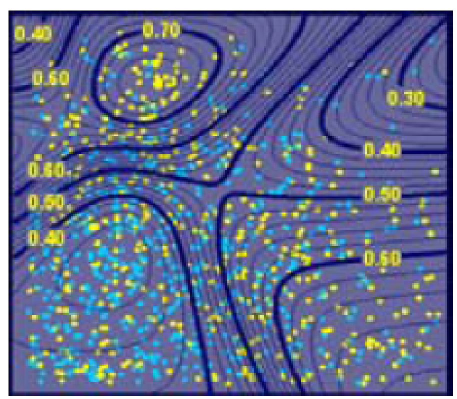


SAS EM Analytic Strengths



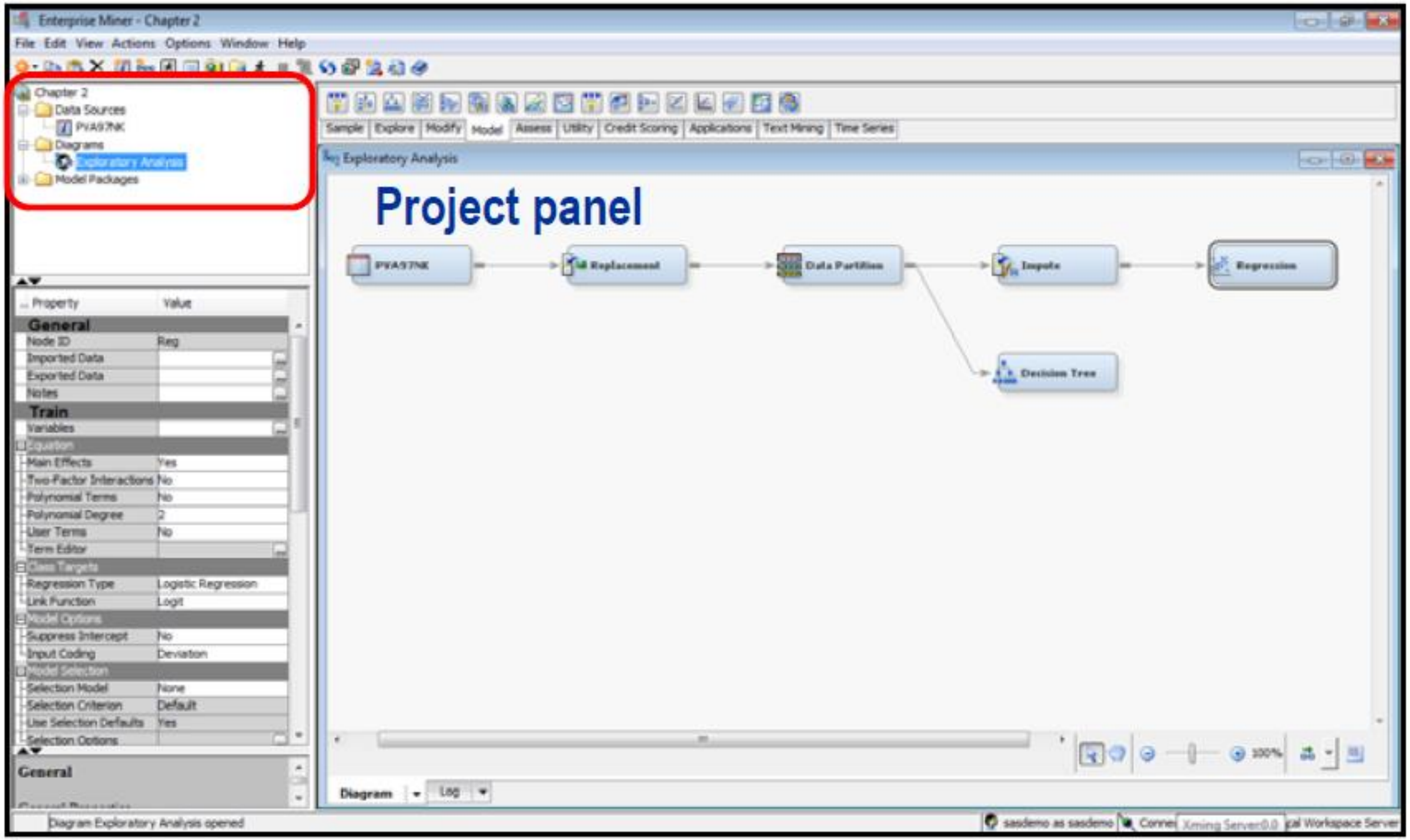


Pattern Discovery

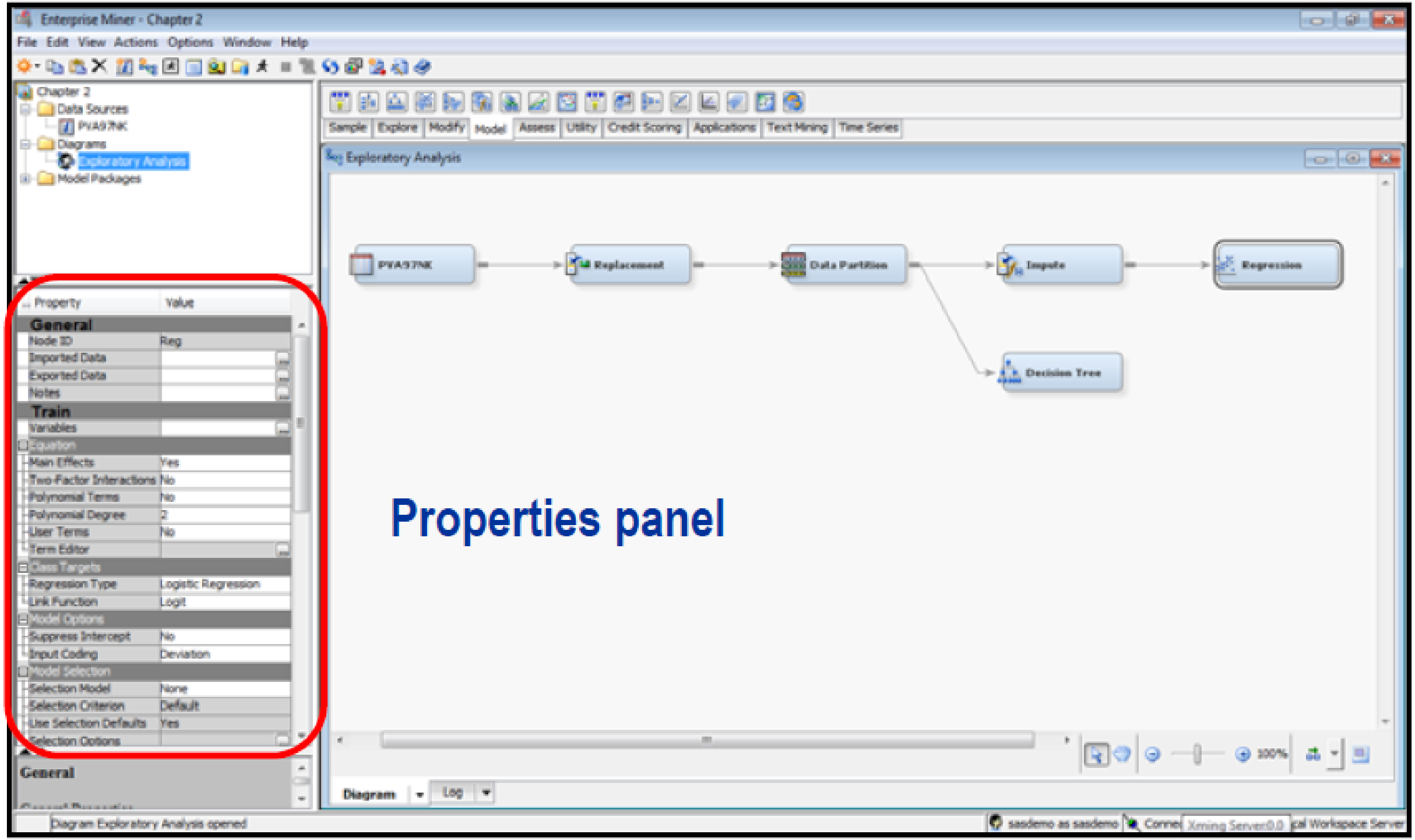


Predictive Modeling

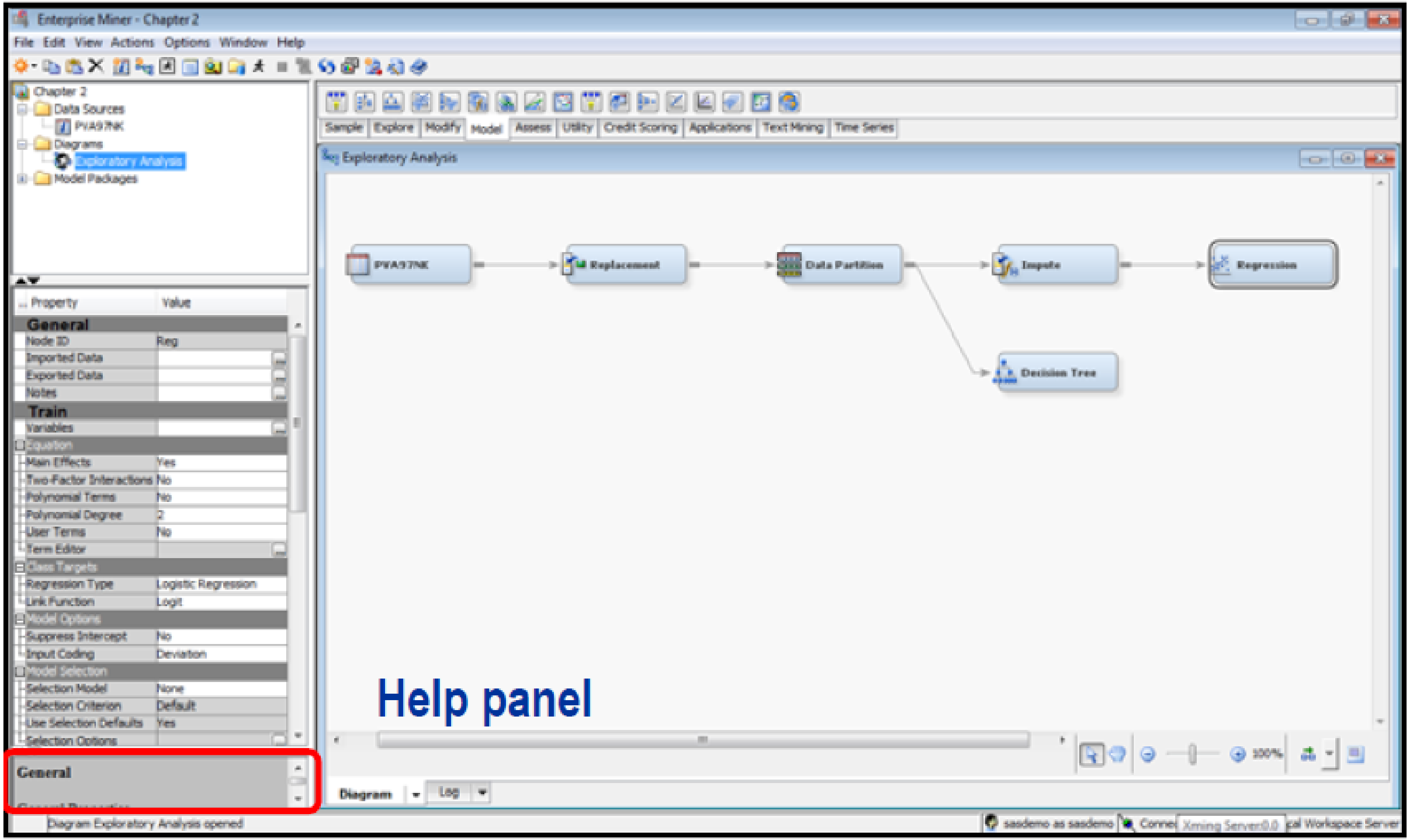
SAS EM – Interface Tour



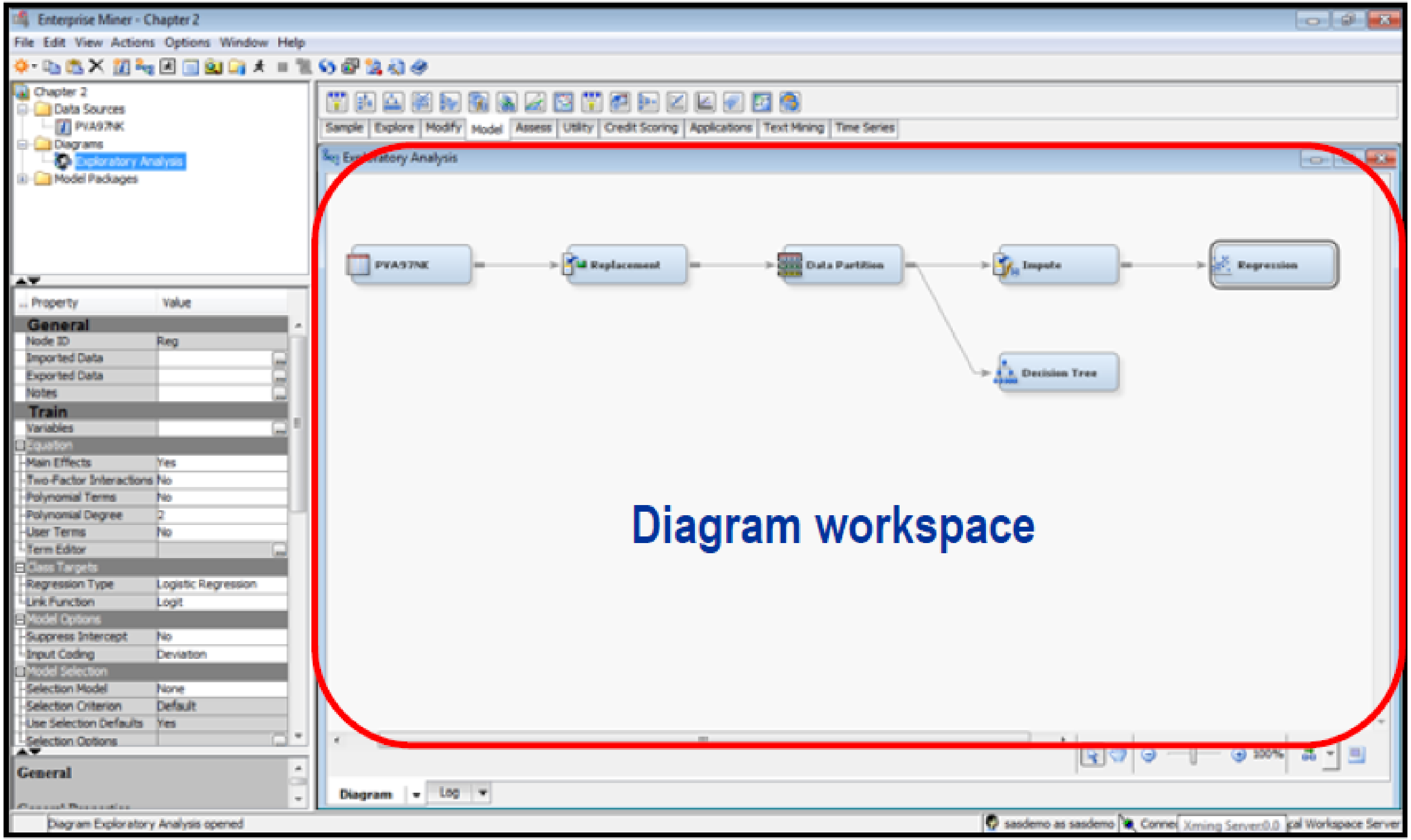
SAS EM – Interface Tour



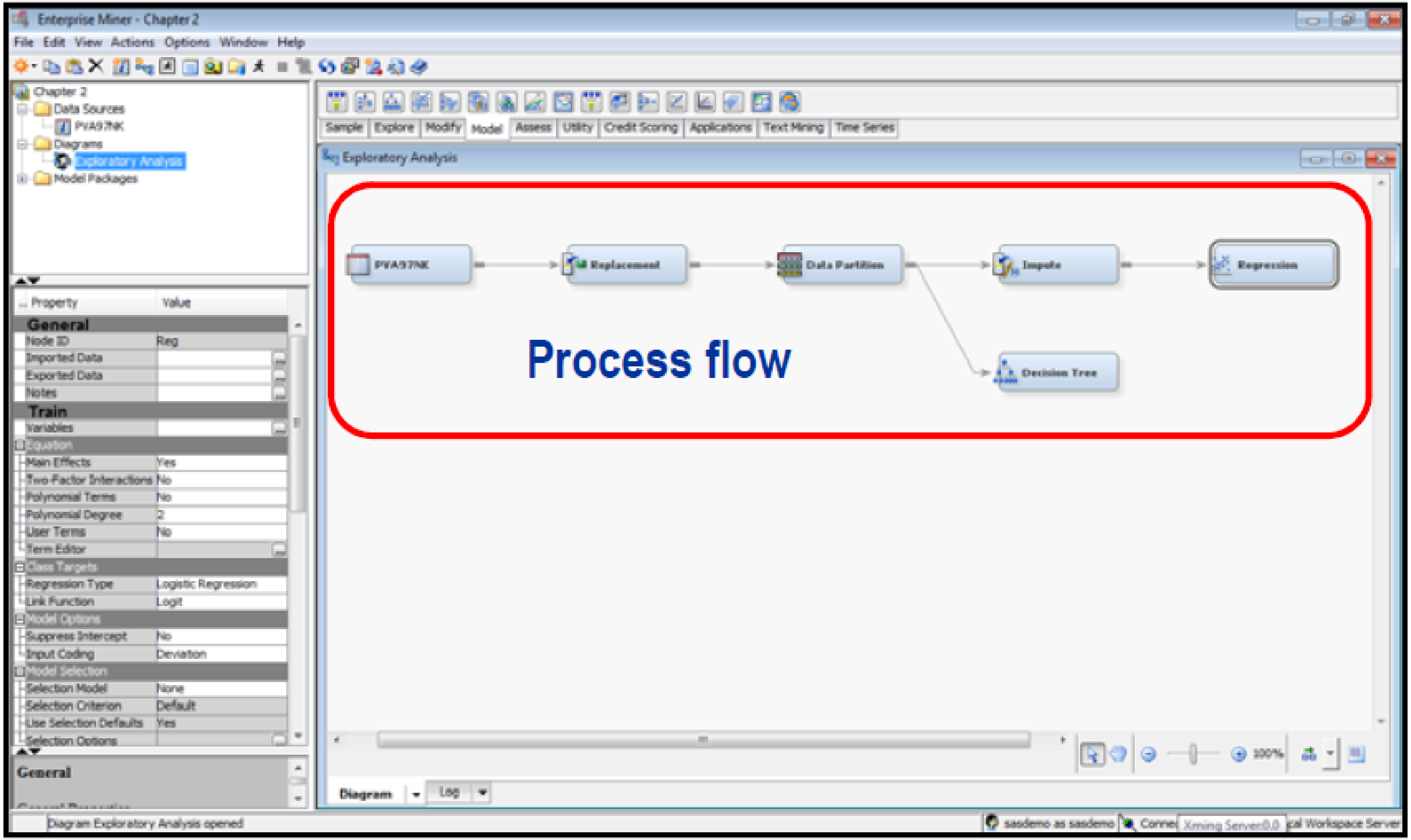
SAS EM – Interface Tour



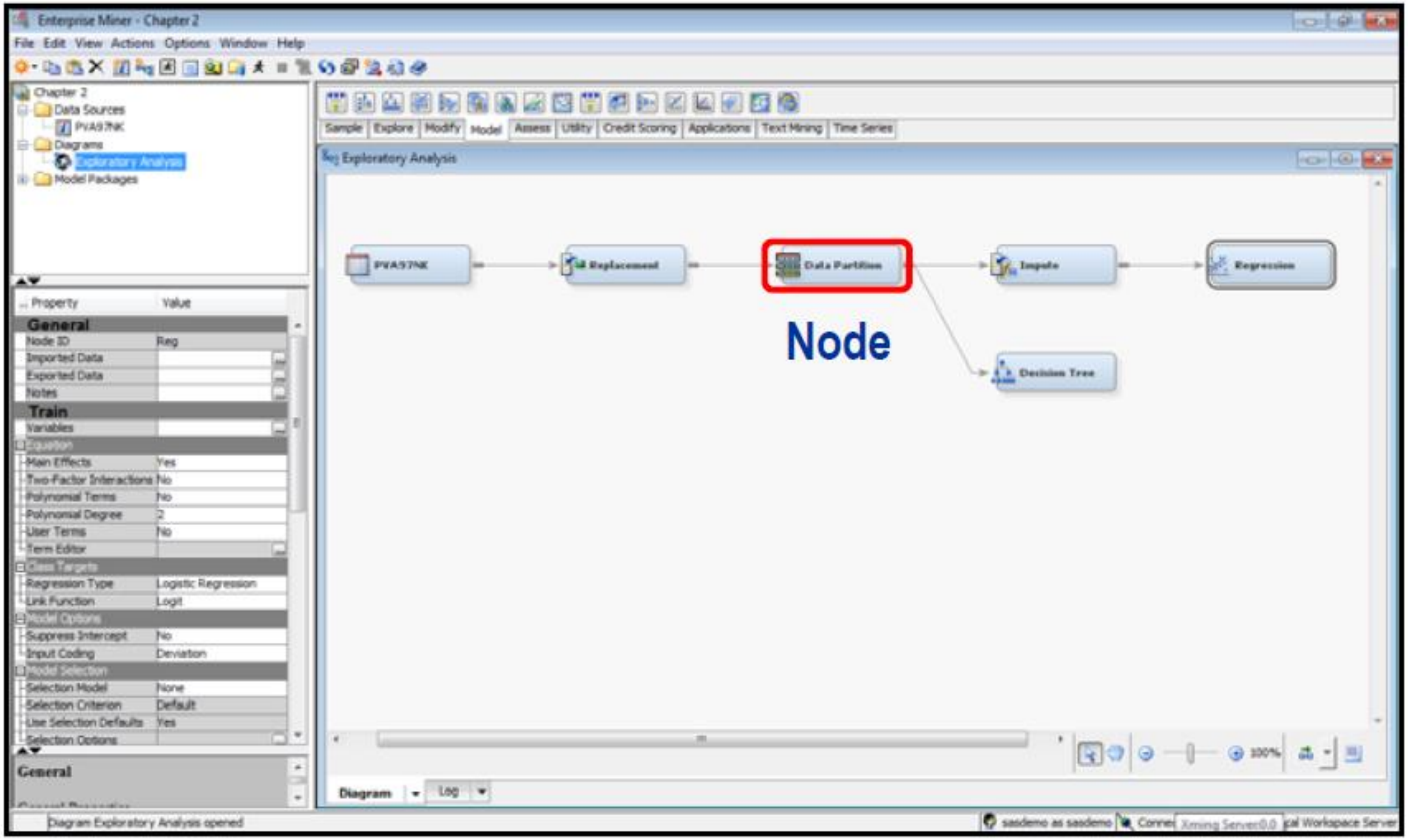
SAS EM – Interface Tour



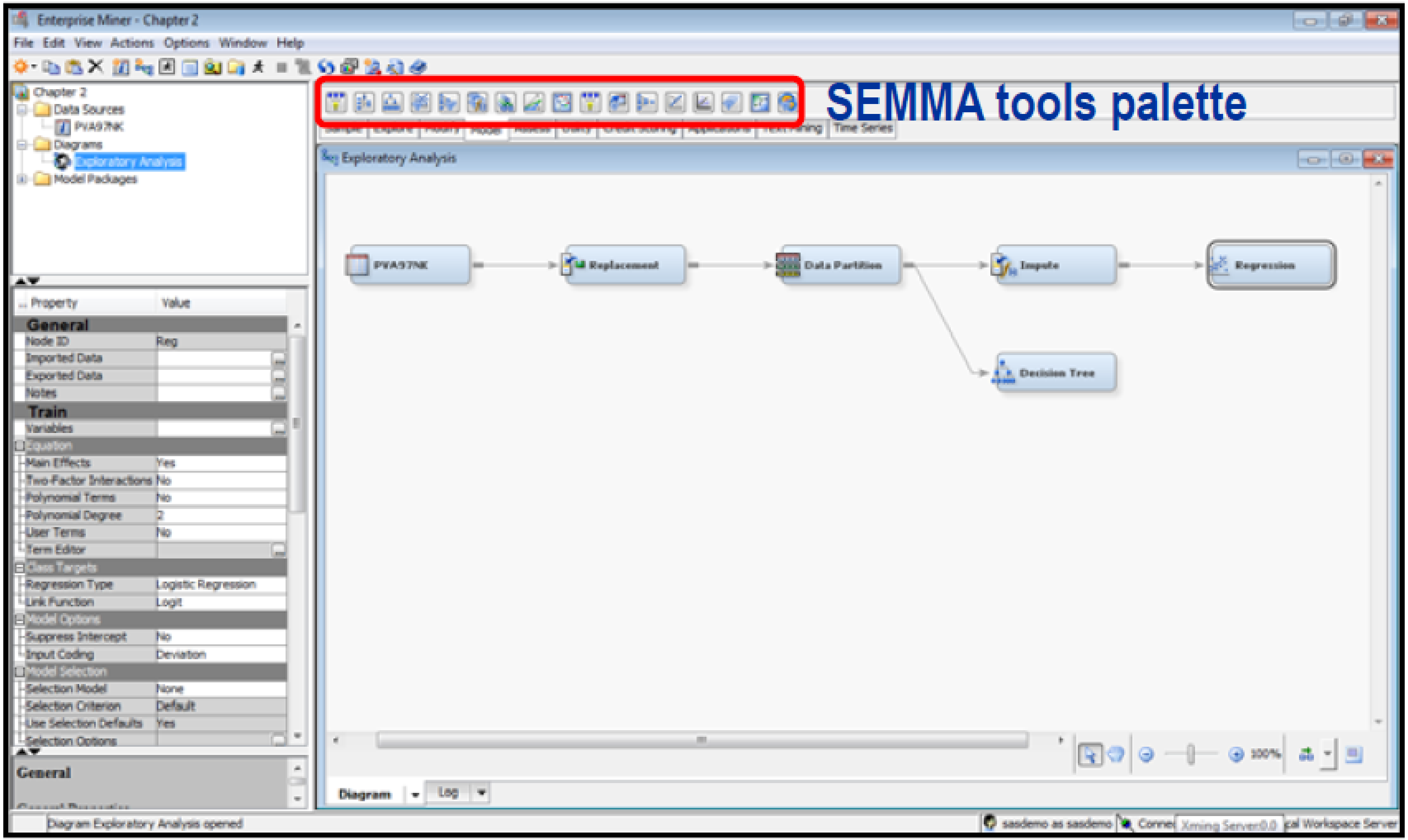
SAS EM – Interface Tour



SAS EM – Interface Tour



SAS EM – Interface Tour



Three Mostly Used Modeling Algorithms

1. Regression

    Models with Binary Target
    Models with an Ordinary Target
    Models with a Nominal (Unordered) Target
    Models with Continuous Target

Models with Binary Target: Logistic Regression

- Since we observe a 0 or a 1, OLS is not an option
- We need a different approach: logistic regression
- The probability of getting 1 depends upon x
- The computation of prob. of event is done through a link function
- $\text{Log}\left[\frac{p\ (y=1/x)}{1-p\ (y=1/x)} \right] = \beta' x$

The linear predictor can be written as: $a+\beta' x$

where x is a vector of inputs and β is the vector of coefficients estimated by Regression Node

Deciding the Best Level of Complexity

- The model with the fewest terms (parsimonious)
- The model with largest (smallest) value of our criteria index (adj. r-square, misclassification rate, AIC, BIC, SBC etc.)
- Using the validation set to compute the criteria (fit index) for each model and then choose the “best”

Fit Indices (Statistics)

- Default
- Akaike's Information Criterion
- Average Squared Error
- Mean Squared Error
- ROC
- Captured Response
- Gain
- Gini Coefficient
- Kolmogorov-Smirnov Statistic
- Lift
- Misclassification Rate
- Average Profit/Loss
- Percent Response
- Cumulative Captured Response
- Cumulative Lift
- Cumulative Percent Response

Three Mostly Used Modeling Algorithms

2.Decision Tree

Very simple to understand

Easy to use

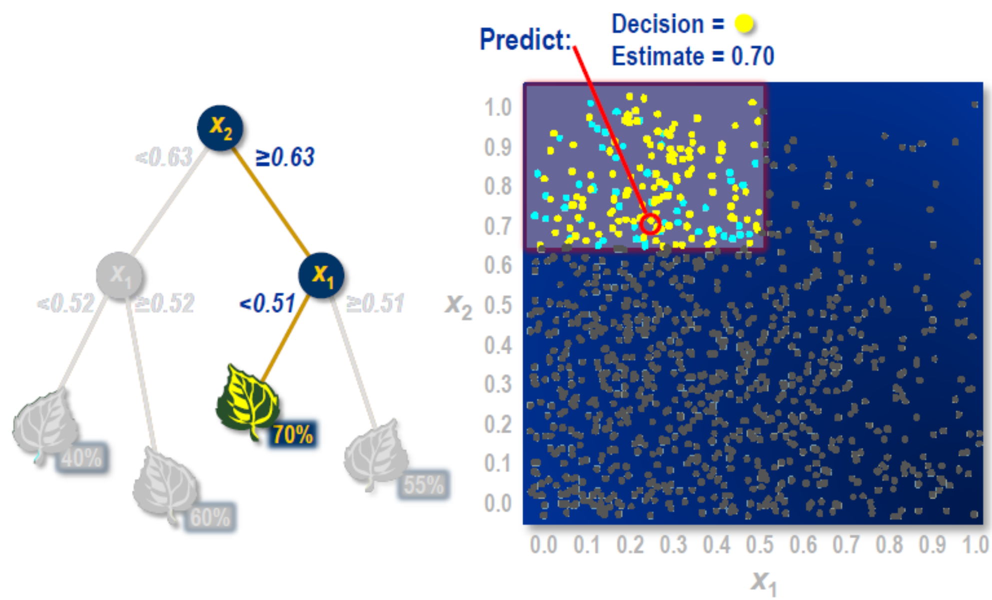
Can explain to the supervisor

Decision Tree Prediction Rules

Chi-Square (Log-worth= -log (χ^2 p-value)

GINI $p_1^2 + p_2^2$

Entropy $(-1 p_1 \log_2(p_1) + p_2 \log_2(p_2))$



Three Mostly Used Modeling Algorithms

2.Neural Net

Very complex mathematical equations

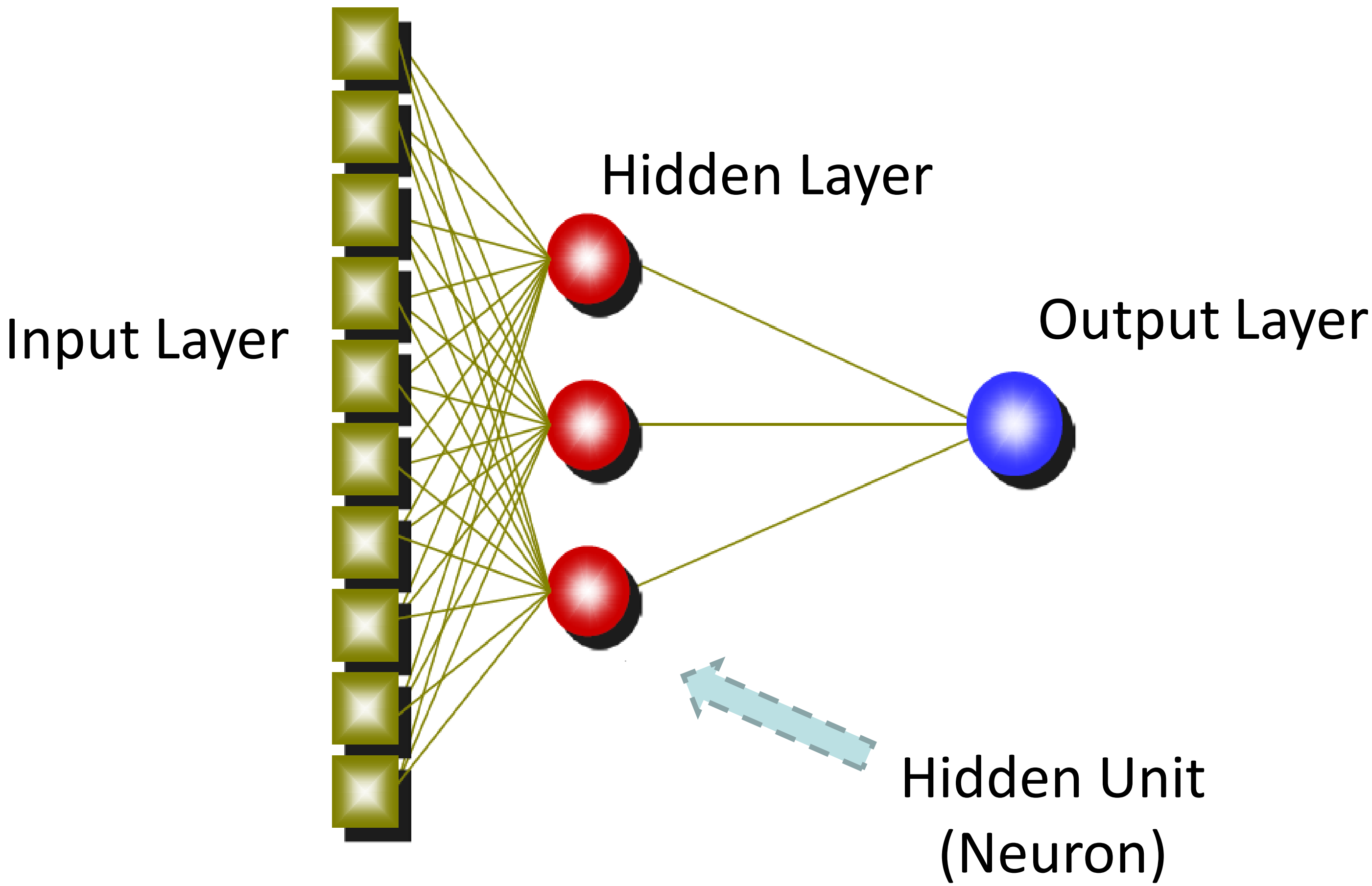
Interpretations of the meaning of the input variables are not possible with final model

Very flexible in accommodating non-linear associations between inputs and target

Two Cultures

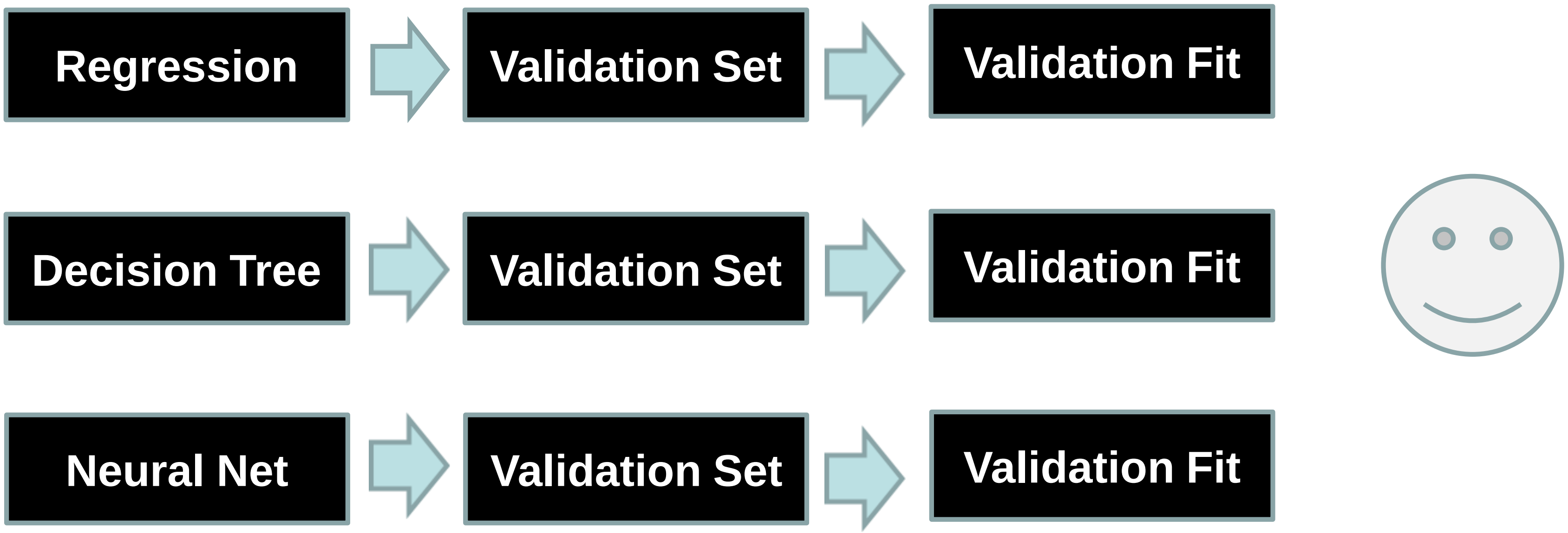
Machine Learning	Statistics
Biological Simulation	Predictive Modeling
Features	Variables
Inputs	Independent Variables
Outputs	Dependent Variables
Synaptic Weights	Parameter Values
Bias	Intercept
Neurons	Terms
Learning	Fitting Models

Multilayer Perceptron (MLP)



Developed with the intention to resemble how the human brain works

Overall Comparison



Find the model of optimal complexity for each family, and then choose overall champion, based on validation performance

Demo of Software

Goal:

To predict as many current 4G customers as possible correctly analyzing existing customer usage and demographic data so that cellular company can identify which customers are likely to switch to 4G network.

Data Description:

A sample dataset of 20,000 3G network customers and 4,000 4G network customers has 249 input variables, one ID variable, and one categorical target variable "Customer_Type" (3G/4G). A 4G customer is defined as a customer who has a 4G Subscriber Identity Module (SIM) card and is currently using a 4G network compatible cellular phone. Three-quarters of the dataset (15,000 3G and 3,000 4G) have the target field and used for model training and validation. The remaining portion of dataset is the scoring data with 5,000 3G and 1,000 4G customers without target variable to test the predictive capability of a developed model.

Some Applications of DM in Health Analytics

- Treatment Effectiveness
- Customer Relationship Management
- Health care management
- Tracking Fee-for-service and Value-based Payer Contracts
- Monitoring and Predicting Fee-for-service Volumes
- Improving Primary Care Reporting
- Predicting Patient Population Risk
- Preventing Hospital Readmissions
- Preventing fraud and abuse

Limitations of Data Mining in Health Analytics

- Accessibility of data
- Missing, corrupted, inconsistent, or non-standardized data
- Fear of data dredging or fishing
- Requiring domain knowledge statistical and research expertise, and IT and data mining knowledge and skills

Future Directions

- Standardization of clinical vocabulary and the sharing of data across organizations to enhance the benefits of healthcare data mining applications.

- Should not be limited to just quantitative data but the use of text mining to be explored.

- There is some progress of using digital diagnostic images in data mining applications.

References

1. Milley, A. (2000). Healthcare and data mining. *Health Management Technology*, 21(8), 44-47.
2. Biafore, S. (1999). Predictive solutions bring more power to decision makers. *Health Management Technology*, 20(10), 12-14.
3. Silver, M. Sakata, T. Su, H.C. Herman, C. Dolins, S.B. & O'Shea, M.J. (2001). Case study: how to apply data mining techniques in a healthcare data
4. Benko, A. & Wilson, B. (2003). Online decision support gives plans an edge. *Managed Healthcare Executive*, 13(5), 20.
5. Kolar, H.R. (2001). Caring for healthcare. *Health Management Technology*, 22(4), 46-47.
6. Relles, D. Ridgeway, G. & Carter, G. (2002). Data mining and the implementation of a prospective payment system for inpatient rehabilitation.
7. *Health Services & Outcomes Research Methodology*, 3(3-4), 247-266.
8. Trybula, W.J. (1997). Data mining and knowledge discovery. *Annual Review of Information Science and Technology*, 32, 197-229.
9. Chung, H.M. & Gray, P. (1999). Data mining. *Journal of Management Information Systems*, 16(1), 11-16.
10. Kreuze, D. (2001). Debugging hospitals. *Technology Review*, 104(2), 32.
11. Veletsos, A. (2003). Getting to the bottom of hospital finances. *Health Management Technology*, 24(8), 30-31.
12. Dakins, D.R. (2001). Center takes data tracking to heart. *Health Data Management*, 9(1), 32-36.
13. Johnson, D.E.L. (2001). Web-based data analysis tools help providers, MCOs contain costs. *Health Care Strategic Management*, 19(4), 16-19.
14. Piazza, P. (2002). Health alerts to fight bioterror. *Security Management*, 46(5), 40.
15. Brewin, B. (2003). New health data net may help in fight against SARS. *Computerworld*, 37(17), 1, 59.
16. Hallick, J.N. (2001), Analytics and the data warehouse. *Health Management Technology*, 22(6), 24-25.
17. Rafalski, E. (2002). Using data mining and data repository methods to identify marketing opportunities in healthcare. *Journal of Consumer Marketing*, 19(7), 607-613.

Q & A

Thank you