Introduction to Data Mining

SAS® Global Forum 2015
Handout

Introduction to Data Mining: SAS® Global Forum 2015 Handout

Prepared date 21APR15

1.1 Introduction to Data Mining

Objectives

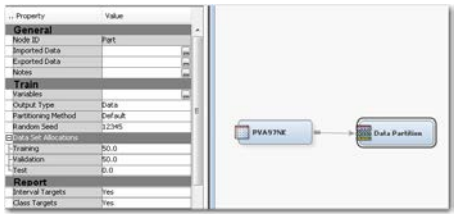
- Overview the main principles and best practices in Data Mining.
- Give a high level overview of three widely used modeling algorithms.

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

Honest Assessment: A Basic principle of Data Mining

- Splitting the data:



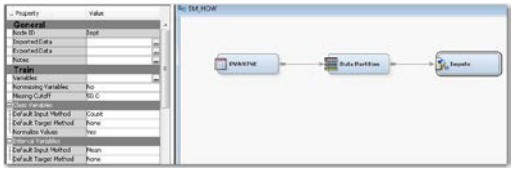
Data Mining: Missing Values

- Decision Trees have built in methods for handling missing values.
- Equation ‘type’ algorithms, e.g. Logistic Regression and Neural Networks, do Complete Case Analysis.
- Imputation is a best practice for equation ‘type’ algorithms.

5

Data Mining: Missing Values

Imputation:



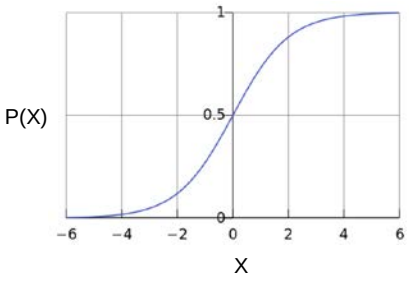
6

Logistic Regression

- Since we observe a 0 or a 1, ordinary least squares is not an option.
- We need a different approach
- The probability of getting a 1 depends upon X.
- We write that as p(X).
- Log odds = log(p(X)/(1-p(X))= a + bX

7

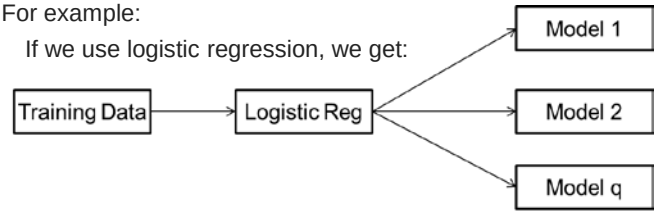
Logistic Graph – Solve for p(X)



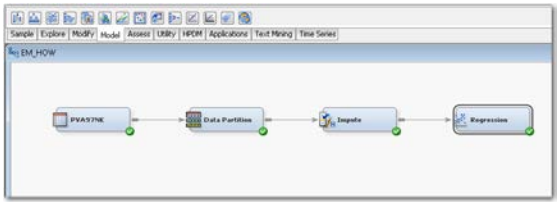
Sequence of Increasingly Complex Models on the Training Set

- For a given procedure (logistic or neural net or decision tree) we use the training set to generate a sequence of models.
- For example:

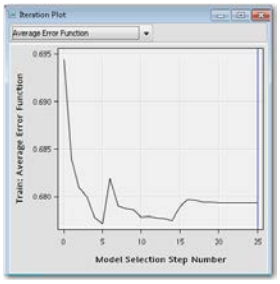
  If we use logistic regression, we get:



Sequence of Increasingly Complex Models on the Training Set



Regression, Ignoring Validation Data



11

How Do We Decide What Level of Complexity is Best?

1) We want the model with the fewest terms (most parsimonious).

2) We want the model with largest (smallest) value of our criteria index (adjusted r-square, misclassification rate, AIC, BIC, SBC etc.)

3) We use the validation set to compute the criteria (Fit Index) for each model and then choose the "best."

12

Fit Indices (Statistics)

- **Default** — The default selection uses different statistics based on the type of target variable and whether a profit/loss matrix has been defined.
  - If a profit/loss matrix is defined for a categorical target, the average profit or average loss is used.
  - If no profit/loss matrix is defined for a categorical target, the misclassification rate is used.
  - If the target variable is interval, the average squared error is used.
- **Akaike's Information Criterion** — chooses the model with the smallest Akaike's Information Criterion value.
- **Average Squared Error** — chooses the model with the smallest average squared error value.
- **Mean Squared Error** — chooses the model with the smallest mean squared error value.
- **ROC** — chooses the model with the greatest area under the ROC curve.
- **Captured Response** — chooses the model with the greatest captured response values using the decile range that is specified in the Selection Depth property.

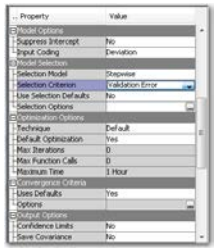
13

Fit Indices (Statistics) – *Continued*

- **Gain** — chooses the model with the greatest gain using the decile range that is specified in the Selection Depth property.
- **Gini Coefficient** — chooses the model with the highest Gini coefficient value.
- **Kolmogorov-Smirnov Statistic** — chooses the model with the highest Kolmogorov - Smirnov statistic value.
- **Lift** — chooses the model with the greatest lift using the decile range that is specified in the Selection Depth property.
- **Misclassification Rate** — chooses the model with the lowest misclassification rate.
- **Average Profit/Loss** — chooses the model with the greatest average profit/loss.
- **Percent Response** — chooses the model with the greatest % response.
- **Cumulative Captured Response** — chooses the model with the greatest cumulative % captured response.
- **Cumulative Lift** — chooses the model with the greatest cumulative lift.
- **Cumulative Percent Response** — chooses the model with the greatest cumulative % response.

14

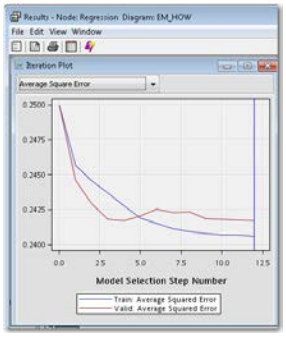
Optimal Complexity



Validation Error will be used to select the Regression model of optimal complexity.

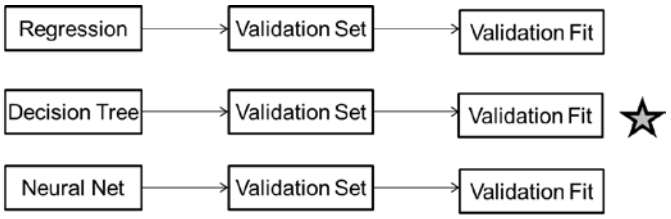
15

Optimal Complexity



16

When More Than One Family of Models is Considered:



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

17

Additional Models

- Decision Tree
- Neural Network

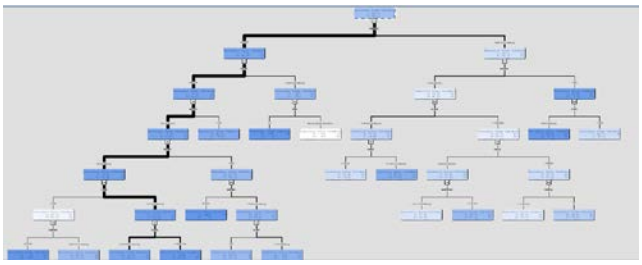
18

Decision Tree

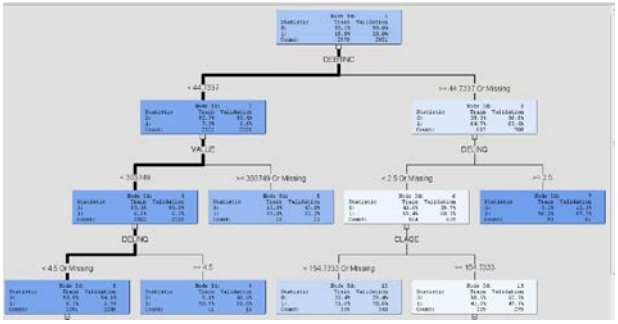
- Very Simple to Understand
- Easy to use
- Can explain to the boss/supervisor

19

Maximal Tree – Ignoring Validation Data



Optimal Tree



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.

Neural Net Diagram

Overall Comparison

Overall Comparison