Data Preparation for Predictive Analytics

Course Outline

I. Context

- When: Which stage(s) of the Modeling Practice Framework involve Data Prep?
- Who: IT, Data Scientists, Domain Experts (or all?)
- Where: Is it taking place in the DW, or on live data using 'real time analytics', maybe in the Data Lake, or the analyst's client-side laptop?
- Why: What needs to be accomplished? What are the assumptions and goals?

II. Stages of Data Preparation (Data Prep 'Theme' of the MPF)

- Data Prep and Project Viability
- Initial Exploration
- Designing the Analytical Sandbox
- Creating the Sandbox
- Data Exploration
- Modeling Prep
- Post Modeling Considerations & Preparing for Deployment

III. Integration

- Team, tools, infrastructure, and ETL: Best Practices
- Merging
- Appending
- Aggregating
- Transposing

IV. Exploration: What is the analyst looking for?

- Analyst collaboration with SMEs
- Types of variables and standard data visualizations

- Examples and/or Demonstration

V. Missing Data, Outliers, and 'Quirks'

- What can you safely get rid of? What will not be deployed?
- What is risky to get rid of?
- Techniques for dealing with missing data
- Techniques for removing noise
- Techniques for fixing problems (including imputation)

VI. Construction

VII. Transformation, Formatting, and 'Conditioning'

VIII. Sampling, Balancing, and Partitioning

IX. Data Prep Considerations for Choosing the Final Model and Deployment