Predictive Analytics & Data Mining: Model Development

Course Summary

Description

The Model Development course presents a deep dive into the data mining process at a tactical level. Attendees will observe demonstrations of machine learning methods and computer-guided analytical techniques for extracting and interpreting complex patterns and relationships from large volumes of data. If you desire an intensive tactical orientation to data mining concepts, tools, techniques and supporting methods, then this event is designed for you.

This course broadly covers data-driven information discovery techniques and model-building tactics without restriction to any particular modeling tool. Popular open-source and commercial packages are leveraged to illustrate methods, but not to showcase the tools.

Objectives

At the end of this course, students will be able to:

- Understand the data mining process and general implementation
- Prepare raw data and benefit from visualization
- Understand various data mining methods and how they compare
- Use advanced model building techniques
- Understand results analysis and validation
- Understand technology and product selection
- Understand solution integration, ongoing performance and maintenance
- Understand where to begin and how to obtain resources and support

Topics

- Introduction
- Crisp-DM Methodology: Parts 3, 4, 5
- Modeling (CRISP 4)
- Evaluation (CRISP 5)
- Wrap Up and Parting Thoughts

Audience

- IT Professionals
- Project Leaders
- Decision Support System Architects
- Business Analysts
- Functional Analysts

Prerequisites

There are no prerequisites for this course. However, participants will benefit by reviewing the CRISP-DM guide ahead of the training.

Duration

Two days

Due to the nature of this material, this document refers to numerous hardware and software products by their trade names. References to other companies and their products are for informational purposes only, and all trademarks are the properties of their respective companies. It is not the intent of ProTech Professional Technical Services, Inc. to use any of these names generically

Predictive Analytics & Data Mining: Model Development

Course Outline

I. Introduction

A. What you will get in this course
B. What is PA/DM?
C. Definition
D. Related terms and fields
E. Machine learning
F. Computer-aided pattern discovery
G. Business analytics and statistics
H. Others you have heard?
I. Examples
J. Differences
K. How can you develop PA/DM opportunities
L. Generative questions
M. Examples
N. Nuts and bolts of a project
O. Big Picture: Introduction to CRISP-DM
P. What is it? What is it not?
Q. Why do we care? Why use it? What is it good for?
R. Example: Tour of CRISP-DM in real-world context
S. Team Exercise
T. One Practitioner's View
U. Regarding PA/DM: What's hype and what isn't?
V. How to be successful with PA/DM
W. Tools and products
X. People matter

II. CRISP-DM Methodology: Parts 3, 4, 5

A. Highlight CRISP-DM 1, 2, 6
B. Business understanding
C. Data understanding
D. Deployment
E. Data Preparation (CRISP 3)
F. Rows: Select data
G. How much data?
H. Rows: Selecting the "unit of analysis"
I. Determine what the record will look like
J. Determine how many records we have to work with
K. Site selection example
L. Rows: Defining the population / outcome of interest
M. Rows: Sampling methods / oversampling
N. Rows: Exclusions / rules of thumb
O. Columns: Identifying types
P. Need definitions (from clients or internal) so that we understand what the data represents. Don't assume that an element isn't important
Q. Categorical / Nominal (what does null mean)?
R. Ordinal
S. Interval / Rational
T. Date / Time
U. Sub-Types (money, count, geo, id, etc, and why care?)
V. Columns: Appropriate statistics and visualizations
W. Univariate
X. Multivariate
Y. Columns: Selection for modeling
Z. See "Clean Data" for pre-modeling elimination of redundant, constant, etc columns
AA. Final selection is done during the Modeling phase
BB. Document the above in a "Scorecard"

III. Modeling (CRISP 4)

A. Select modeling technique
B. Taxonomies: An overview
C. Supervised vs. Unsupervised
D. Descriptive vs. Predictive
E. Classification vs. Estimation
F. Supervised -- Constellation of methods with pros and cons
G. Classification
H. Decision Trees
I. Logistic Regression
J. Neural Networks
K. K-Nearest Neighbor
L. Prediction
M. Linear Regression
N. Neural Networks
O. MARS
P. Exercise: Scenario revisited -- What method(s) do we choose?
Q. Unsupervised -- More methods with pros and cons
R. Segmentation / Clustering

Due to the nature of this material, this document refers to numerous hardware and software products by their trade names. References to other companies and their products are for informational purposes only, and all trademarks are the properties of their respective companies. It is not the intent of ProTech Professional Technical Services, Inc. to use any of these names generically

Predictive Analytics & Data Mining: Model Development

Course Outline (cont'd)

S. Hierarchical clustering
T. K-Means
U. Decision trees
V. Association Rules
W. Team Exercise: Com up with an expert-derived decision tree to make a selection for supervised problems
X. Advanced Topics
Y. Ensemble / Hybrid Models
Z. Bagging
AA. Boosting
BB. Parting remarks
CC. Models should be as simple as possible, but no simpler
DD. Why not both? (a low-res descriptive model *and*
a high-res opaque accuracy model)
EE. Generate test design
FF. Data segregation
GG. Performance metrics: Whenever possible, go for the custom metric -- "If you build it, they will come."
HH. Build Model
II. Use a tool, select a method, set parameters (if any), and select candidate columns, select outcome (if supervised)
JJ. Variable selection techniques for supervised methods
KK. Variable selection techniques for unsupervised methods
LL. Assess Model (Tweaking)
MM. Predictors
NN. Manually removing or limiting
OO. Forcing predictors
PP. Structure
QQ. Profiles
RR. Compared to What?
SS. Baseline model comparison
TT. Train/Test/Validation comparison
UU. Scoring the model
VV. What does scoring mean?
WW. How is it different from building the model?
XX. What are we looking for when scoring?
YY. Final Product
ZZ. Model(s)
AAA. Description(s)

IV. Evaluation (CRISP 5)

A. Evaluate results (from business perspective)
B. Prelude to business use presentation
C. Informal, low-risk setting
D. Poke holes early, before business presentation
E. Does the model or segmentation make sense?
F. Does it contradict of reinforce the standard "lore"?
G. SWOT analysis: What are the strengths, weaknesses, opportunities and threats?
H. Get support and buy-in from potential champions
I. Candidate names for segments
J. Present results to business users or clients
K. BUs need to be convinced: Models, segments and analysis need to be marketed!
L. Deployment will require change
M. To processes
N. To systems
O. To ingrained mindsets
P. Deployment costs (to each change area above)
Q. Results must have business value, not technical representations
R. Performance results -- in business terms
S. Descriptions
T. No equations
U. Tell the story, paint the vision, what will life be like with or without the model in place?
V. Review Process
W. Follow-ups to the presentation
X. Anticipate follow-up issues in planning and estimates
Y. Revisions to the model(s) or segments based upon feedback
Z. Final quality assurance
AA. Determine next steps
BB. Are you done?

Due to the nature of this material, this document refers to numerous hardware and software products by their trade names. References to other companies and their products are for informational purposes only, and all trademarks are the properties of their respective companies. It is not the intent of ProTech Professional Technical Services, Inc. to use any of these names generically

Predictive Analytics & Data Mining: Model Development

Course Outline (cont'd)

CC. Will the model(s) be deployed? Why or why not?
DD. Document!
EE. Lessons learned meeting
FF. Final Product
GG. Consulting Exercise

V. Wrap-Up and Parting Thoughts

A. Final Q&A
B. Springboard exercise: "The boss in the elevator"
C. PA/DM Philosophy
D. Understand the problem
E. Understand the data
F. Then, think about how to solve it (Einstein quotes)
G. Work on problems with specific business goals, specific hypotheses to be tested. Do NOT go prospecting for "data mining nuggets."
H. Next Steps
I. PA/DM Level II Course: "Strategic Implementation"
J. Certification Exam (for those who complete the series)
K. Product training courses
L. Keep learning!
M. Supplementary materials and resources
N. Conferences and communities
O. Get started on a project!

Due to the nature of this material, this document refers to numerous hardware and software products by their trade names. References to other companies and their products are for informational purposes only, and all trademarks are the properties of their respective companies. It is not the intent of ProTech Professional Technical Services, Inc. to use any of these names generically