

Data Preparation for Data Mining

This Level IV course is intended for data mining and IT professionals interested in transforming raw data into meaningful inputs for predictive models.

Duration: 3.0 days

Course Description [\[Click to register ONLINE \]](#)

Data preparation is universally held as the key to successful data mining. This course introduces programming techniques used by analysts to transform raw data into a form suitable for predictive modeling. The course teaches you how to extract appropriate data from raw data sources and transform transactions or event data to a form that predictive models can utilize. You also learn how to effectively incorporate non-numeric data in predictive models as well as manage exceptional and extreme data. After completing this course, you also will be able to document the data preparation process.

Prerequisites

This course assumes some experience in both data mining and SAS programming. Before attending this course, you should

- have experience with common predictive modeling techniques, which you can gain from the [Predictive Modeling Using SAS Enterprise Miner 5.1](#) course
- have experience with creating, managing, and manipulating SAS data sets, which you can gain from the [SAS Programming I: Essentials](#) and [SAS Programming II: Manipulating Data with the DATA Step](#) courses.

Course Contents

Introduction

- raw data structures
- predictive modeling data structure
- over view of data preparation challenges

Extracting Relevant Data

- data difficulties
- assessing available data
- accessing available data
- drawing a representative target sample
- drawing an uncontaminated input sample

Transforming Transactions or Event Data

- advantages and disadvantages of transactions data
- common transaction structures
- defining the time horizon
- fixed and variable time horizon methods
- implementing common transaction transformations

Using Non-Numeric Data

- definitions and difficulties of non-numeric data
- miscoding and multicoding detection
- controlling degrees of freedom
- geocoding

Managing Exceptions and Extremes

- difficulties with outliers, missing and non-applicable values, and extreme distributions
- detection of exceptions and extremes
- remedies for exceptional and extreme values

Course Materials

Students attend classroom courses in one of our public training centers. You receive a hardcopy of the course notes.