

Supervised Training Data Set

Predictive modeling is designed to describe or predict one or more variables based on other variables in the data; it is also called *supervised training*. Predictive models such as traditional linear regression modeling, decision tree modeling, neural network modeling, nearest neighbor modeling, two-stage modeling, and discriminate analysis may be applied in Enterprise Miner. The main idea in predictive modeling is to either minimize the error or maximize the expected profit. In Enterprise Miner, the modeling terms are distinguished by their model roles. For example, the output response variable that you want to predict would be set to a **target** model role and all the predictor variables in the predictive model are assigned a model role of **input**. Time identifier or carryover variables might be passed along in the modeling design that identifies each observation in the data with an **id** model role.

The following is the data set that is used in the various Enterprise Miner nodes based on supervised training in explaining both the configuration settings and the corresponding results. The SAS data set is called HMEQ. The data set is located in the SAMPSIO directory within the folder in which your SAS software is installed. The SAMPSIO directory is automatically available for access once Enterprise Miner is opened. The data consists of applicants granted credit for a certain home equity loan that has 5,960 observations. The categorical target variable that was used in the following examples is a binary-valued variable called BAD that identifies if a client either defaulted or repaid their home equity loan. For interval-valued targets, the variable called DEBTINC, which is the ratio between debt and income, was used in many of the following modeling examples. There are thirteen variables in the data mining data set with nine numeric variables and four categorical variables. The following table displays the variables in the data set, the model role, measurement level, and variable description. The database was used in many of the following predictive modeling designs to determine if the applicant can be approved for a home equity loan.

| Name | Model Role | Measurement Level | Description |
|---------|------------|-------------------|---|
| BAD | Target | Binary | 1 = Defaulting on the loan, 0 = Repaid the loan |
| CLAGE | Input | Interval | Age (in months) of the oldest trade line |
| CLNO | Input | Interval | Number of trade lines |
| DEBTINC | Input | Interval | Ratio of debt to income. |
| DELINQ | Input | Interval | Number of delinquent trade lines |
| DEROG | Input | Interval | Number of major derogatory reports |
| JOB | Input | Nominal | Six occupational categories |
| LOAN | Input | Binary | Amount of the loan request |
| MORTDUE | Input | Interval | Amount due on the existing mortgage |
| NINQ | Input | Interval | Number of recent credit inquires |
| REASON | Input | Binary | DebtCon = debt consolidation, HomeImp = home improvement |
| VALUE | Input | Interval | Current property value |
| YOJ | Input | Interval | Number of years at the present job |