Data sources

You or the administrator need to specify the data sources to use in the IBM® Predictive Customer Intelligence solution for modeling, analysis, simulation and testing, and scoring.

Use IBM Analytical Decision Management to plan the model and to decide which data sources to use.

You need the following types of data in the modeling process:

**Historical or analytical data**

To build the model, you need information about what to predict. For example, if you want to predict churn, you need information about customers such as their complaints history, number of months since they upgraded their plan, sentiment score, demographic history, and estimated income. This is often referred to as historical data or analytical data, and it must contain some or all of the fields in the project data model, plus an additional field that records the outcome or result that you want to predict. This extra field is used as the target for modeling.

**Operational or scoring data**

To use the model to predict future results, you need data about the group or population that you are interested in, such as incoming claims. This is often referred to as operational data or scoring data. The project data model is typically based on this data.

You can use the following types of data sources:

* A database that supports ODBC, such as IBM DB2®.
* An Enterprise View that is defined in IBM SPSS® Collaboration and Deployment Services.
* A file that is used by IBM SPSS Statistics, such as a text file (txt), or a comma separated file (csv).

When you add a new data source, map all of its fields to ensure compatibility with the project data model. For example, if the project data model requires a field named purchase with values Yes and No for the measurement level flag, then any data source that you use must have a compatible field. If the field names are not identical, they can be mapped accordingly.

Note that the input and associated mapped field must have the same data type.

You can characterize the information that each data field represents. Define a measurement level to determine how a given field is used in business rules, models, or other applications.

You can derive additional fields or attributes for the application by using the expression manager. For example, if you use banking data, you may want to create an expression that shows the ratio between a customer’s income and the number of loan accounts that the customer has. Expressions are always numeric with a measurement of Continuous; this cannot be changed.

To enforce corporate-wide policies, use global selections to choose the records to include or exclude from processing by the application. For example, you might have a corporate-wide policy to exclude customers with poor credit or payment histories from future mortgage campaigns. Global selections can be particularly effective when used in combination with shared rules. Shared rules are saved as separate objects that can be used by multiple applications. If the shared rule changes, all applications that use the rule can then be updated.

Data mining that uses IBM SPSS Modeler focuses on the process of running data through a series of nodes. This is referred to as a stream. This series of nodes represents operations to be performed on the data, while links between the nodes indicate the direction of data flow. Typically, you use a data stream to read data into IBM SPSS Modeler, run it through a series of manipulations, and then send it to a destination, such as a table or a viewer.

For example, to open a data source, you add a new field, select records based on values in the new field, and then display the results in a table. In this example, your data stream would consist of the following nodes:

* Variable File node, which reads the data from the data source
* Derive node, which adds the new calculated field to the data set
* Select node, which uses the selection criteria to exclude records from the data stream
* Table node, which displays the results of your manipulations on screen