Classification, Regression, Segmentation and Clustering Scenarios
Automated Analytics User Guide
# Content

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>What's New in Classification, Regression, Segmentation and Clustering Scenarios.</td>
<td>8</td>
</tr>
<tr>
<td>1.1</td>
<td>Document History.</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>Welcome to this Guide.</td>
<td>9</td>
</tr>
<tr>
<td>2.1</td>
<td>About this Document.</td>
<td>9</td>
</tr>
<tr>
<td>2.2</td>
<td>Which Sections should you Read?.</td>
<td>9</td>
</tr>
<tr>
<td>2.3</td>
<td>Files and Documentation Provided with this Guide.</td>
<td>11</td>
</tr>
<tr>
<td>3</td>
<td>Overview of Automated Analytics</td>
<td>12</td>
</tr>
<tr>
<td>3.1</td>
<td>Automated Analytics Architecture.</td>
<td>12</td>
</tr>
<tr>
<td>3.2</td>
<td>Automated Analytics Operational Phases.</td>
<td>13</td>
</tr>
<tr>
<td>3.3</td>
<td>Methodological Prerequisites.</td>
<td>15</td>
</tr>
<tr>
<td>4</td>
<td>Essential Concepts.</td>
<td>17</td>
</tr>
<tr>
<td>4.1</td>
<td>Operation of Automated Analytics: Overview.</td>
<td>17</td>
</tr>
<tr>
<td>4.2</td>
<td>Data Sources Supported.</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>SAP HANA as a Data Source.</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>Managing Performance when Using Databases.</td>
<td>20</td>
</tr>
<tr>
<td>4.3</td>
<td>Training and Application Dataset.</td>
<td>21</td>
</tr>
<tr>
<td>4.4</td>
<td>Cutting Strategies.</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>Customized Cutting Strategy.</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>Automatic Cutting Strategies.</td>
<td>23</td>
</tr>
<tr>
<td>4.5</td>
<td>Table of Data.</td>
<td>26</td>
</tr>
<tr>
<td>4.6</td>
<td>Variables.</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>Continuous Variables.</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>Ordinal Variables.</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>Nominal Variables.</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>Variable Storage Formats.</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Target Variable.</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>Explanatory Variable.</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>Weight Variable.</td>
<td>34</td>
</tr>
<tr>
<td>4.7</td>
<td>Models.</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>Generating a Model.</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>Representation of a Model.</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>Validating the Model.</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td>How to Obtain a Better Model.</td>
<td>38</td>
</tr>
<tr>
<td>4.8</td>
<td>Performance Indicators.</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>Quality Indicator: Predictive Power.</td>
<td>39</td>
</tr>
</tbody>
</table>
Step 1 - Defining the Modeling Parameters ........................................ 166
Step 2 - Generating and Validating the Model ................................ 176
Step 3 - Analyzing and Understanding the Model Generated ............ 179
Step 4 - Using the Model .............................................................. 205

7 Glossary .................................................................................. 214
7.1 analytical data management .......................................................... 214
7.2 analytical dataset ................................................................. 214
7.3 analytical record ................................................................ 214
7.4 antecedent ................................................................. 214
7.5 application dataset ................................................................. 215
7.6 association rule ............................................................... 215
7.7 attribute ................................................................. 215
7.8 AUC ................................................................. 215
7.9 authenticated server ............................................................... 215
7.10 autoselection ................................................................. 216
7.11 bin ................................................................. 216
7.12 bipartite graph display / non-bipartite graph display ................. 216
7.13 bubble chart ............................................................... 216
7.14 calendar table ............................................................... 216
7.15 category ................................................................. 217
7.16 category significance ............................................................. 217
7.17 centroid ................................................................. 217
7.18 chunk (by chunk) ............................................................... 217
7.19 classification rate .............................................................. 217
7.20 confidence ................................................................. 217
7.21 confusion matrix .............................................................. 218
7.22 consequent ................................................................. 218
7.23 continuous variable ............................................................. 218
7.24 contribution ............................................................... 218
7.25 correlation ................................................................. 218
7.26 cross statistics ................................................................. 219
7.27 customized cutting strategy ................................................ 219
7.28 customized profit ............................................................. 219
7.29 cutting strategy ............................................................... 219
7.30 data aggregation .............................................................. 219
7.31 dataset ................................................................. 220
7.32 data source ................................................................. 220
7.33 database ................................................................. 220
7.34 descriptive model .............................................................. 220
7.35 detected profit .............................................................. 220
7.36 determination coefficient (R2) ............................................... 220
### Classification, Regression, Segmentation and Clustering Scenarios

<table>
<thead>
<tr>
<th>Page</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.37</td>
<td>deviation.</td>
</tr>
<tr>
<td>7.38</td>
<td>domain.</td>
</tr>
<tr>
<td>7.39</td>
<td>encoding.</td>
</tr>
<tr>
<td>7.40</td>
<td>engine.</td>
</tr>
<tr>
<td>7.41</td>
<td>entity.</td>
</tr>
<tr>
<td>7.42</td>
<td>error bar.</td>
</tr>
<tr>
<td>7.43</td>
<td>error mean.</td>
</tr>
<tr>
<td>7.44</td>
<td>error standard deviation.</td>
</tr>
<tr>
<td>7.45</td>
<td>event dataset.</td>
</tr>
<tr>
<td>7.46</td>
<td>excluded variable.</td>
</tr>
<tr>
<td>7.47</td>
<td>explanatory variable.</td>
</tr>
<tr>
<td>7.48</td>
<td>Expression Editor.</td>
</tr>
<tr>
<td>7.49</td>
<td>extra-predictable variable.</td>
</tr>
<tr>
<td>7.50</td>
<td>false positive.</td>
</tr>
<tr>
<td>7.51</td>
<td>fluctuation.</td>
</tr>
<tr>
<td>7.52</td>
<td>GINI index.</td>
</tr>
<tr>
<td>7.53</td>
<td>horizon-wide MAPE.</td>
</tr>
<tr>
<td>7.54</td>
<td>In-database apply.</td>
</tr>
<tr>
<td>7.55</td>
<td>item.</td>
</tr>
<tr>
<td>7.56</td>
<td>itemset.</td>
</tr>
<tr>
<td>7.57</td>
<td>iteration.</td>
</tr>
<tr>
<td>7.58</td>
<td>KL (Kullback-Leibler).</td>
</tr>
<tr>
<td>7.59</td>
<td>KPI (Key Performance Indicator).</td>
</tr>
<tr>
<td>7.60</td>
<td>K-S test.</td>
</tr>
<tr>
<td>7.61</td>
<td>Lift.</td>
</tr>
<tr>
<td>7.62</td>
<td>Lift profit.</td>
</tr>
<tr>
<td>7.63</td>
<td>lower bound.</td>
</tr>
<tr>
<td>7.64</td>
<td>maximum error (LInf).</td>
</tr>
<tr>
<td>7.65</td>
<td>mean.</td>
</tr>
<tr>
<td>7.66</td>
<td>mean absolute error (L1).</td>
</tr>
<tr>
<td>7.67</td>
<td>mean absolute percentage error (MAPE).</td>
</tr>
<tr>
<td>7.68</td>
<td>mean square error (L2).</td>
</tr>
<tr>
<td>7.69</td>
<td>metadata.</td>
</tr>
<tr>
<td>7.70</td>
<td>meta-operator.</td>
</tr>
<tr>
<td>7.71</td>
<td>missing value.</td>
</tr>
<tr>
<td>7.72</td>
<td>monotonicity.</td>
</tr>
<tr>
<td>7.73</td>
<td>multiple-instance installation.</td>
</tr>
<tr>
<td>7.74</td>
<td>nominal variable.</td>
</tr>
<tr>
<td>7.75</td>
<td>normalize.</td>
</tr>
<tr>
<td>7.76</td>
<td>numeric filter.</td>
</tr>
<tr>
<td>7.77</td>
<td>ordinal variable.</td>
</tr>
</tbody>
</table>
outlier. ................................................................. 228
performance indicator (PI). ............................... 228
periodic cutting strategy. ........................................ 229
pivot. .................................................................... 229
polynomial. .............................................................. 229
population. .............................................................. 229
prediction range. .................................................... 229
predictive model. .................................................... 230
profit type. .............................................................. 230
quality indicator: the predictive power (KI). .......... 230
random cutting strategy. ......................................... 230
record. .................................................................. 230
robustness. ............................................................... 231
robustness indicator: the prediction confidence (KR). 231
ROC. ................................................................. 231
role. ..................................................................... 231
root. .................................................................... 231
score. .................................................................... 232
scorecard. ............................................................... 232
seasonal. ................................................................. 232
sensitivity. ............................................................... 232
sequential cutting strategy. ................................. 232
session. ................................................................. 233
simulation. ............................................................. 233
smart variable contributions. ............................... 233
social network analysis. ......................................... 233
standard deviation. ............................................... 233
standardized profit. ............................................... 233
statistical report. .................................................... 234
storage. ................................................................. 234
sub-sampling. ........................................................ 234
support. ................................................................. 234
table of data. ........................................................ 234
target key. .............................................................. 235
target key. .............................................................. 235
target variable. ...................................................... 235
temporal analytical dataset. ................................. 235
time series. ............................................................. 235
timeout. ................................................................. 236
time-stamped population. .................................... 236
training. ................................................................. 236
<table>
<thead>
<tr>
<th>Page</th>
<th>Term</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.119</td>
<td>training dataset</td>
<td>236</td>
</tr>
<tr>
<td>7.120</td>
<td>transaction</td>
<td>236</td>
</tr>
<tr>
<td>7.121</td>
<td>true negative</td>
<td>237</td>
</tr>
<tr>
<td>7.122</td>
<td>true positive</td>
<td>237</td>
</tr>
<tr>
<td>7.123</td>
<td>unassigned record</td>
<td>237</td>
</tr>
<tr>
<td>7.124</td>
<td>upper bound</td>
<td>237</td>
</tr>
<tr>
<td>7.125</td>
<td>variable</td>
<td>237</td>
</tr>
<tr>
<td>7.126</td>
<td>variable pool</td>
<td>237</td>
</tr>
<tr>
<td>7.127</td>
<td>variable type</td>
<td>238</td>
</tr>
<tr>
<td>7.128</td>
<td>weight variable</td>
<td>238</td>
</tr>
</tbody>
</table>
1 What's New in Classification, Regression, Segmentation and Clustering Scenarios

Links to information about the new features and documentation changes for Classification, Regression, Segmentation and Clustering Scenarios.

<table>
<thead>
<tr>
<th>What's New</th>
<th>Link to More Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAP BusinessObjects Predictive Analytics 3.0 now supports</td>
<td>SAP HANA as a Data Source [page 18]</td>
</tr>
<tr>
<td>SAP HANA Attribute Views.</td>
<td></td>
</tr>
<tr>
<td>The Cluster Summary view displays an additional cluster</td>
<td>Cluster Summary [page 193]</td>
</tr>
<tr>
<td>corresponding to the unassigned observations.</td>
<td></td>
</tr>
<tr>
<td>A new code type CCL is available.</td>
<td>List of Generated Codes [page 154]</td>
</tr>
<tr>
<td>Update of the description of the group variables.</td>
<td>How to Describe Selected Variables [page 68]</td>
</tr>
<tr>
<td>Added a graphic displaying the predictive power and the prediction</td>
<td>Predictive Power, Prediction Confidence and Model Graphs</td>
</tr>
<tr>
<td>confidence</td>
<td>[page 40]</td>
</tr>
<tr>
<td>The formulat to calculate the Normal profit is updated.</td>
<td>Category Importance Definition [page 114]</td>
</tr>
</tbody>
</table>

1.1 Document History

<table>
<thead>
<tr>
<th>Product Version</th>
<th>What's Changed</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAP BusinessObjects Predictive Analytics 2.5</td>
<td>Update of the confusion matrix table and explanation.</td>
</tr>
<tr>
<td></td>
<td>See Confusion Matrix [page 123]</td>
</tr>
</tbody>
</table>
2 Welcome to this Guide

2.1 About this Document

This document is addressed to people who want to evaluate or use Modeler to build Regression/Classification and Segmentation/Clustering models. Use of this guide does not require any prior expertise in statistics or databases.

This document introduces you to the basic concepts and main functionalities of Modeler. Using two application scenarios, you can create your first models with confidence. This document is the primary guide to the two application features described in the following table:

<table>
<thead>
<tr>
<th>Feature</th>
<th>Purpose</th>
<th>Example</th>
</tr>
</thead>
</table>
| Regression/Classification| To understand and predict a phenomenon.      | You work for an automobile manufacturer and wish to send a promotional mailing to your prospects. Modeler - Regression/Classification allows you to:  
  - Understand why previous prospects responded to such a mailing.  
  - Predict the response rate to such a mailing sent to new prospects. |
| Segmentation/Clustering  | To describe a dataset, by breaking it down into homogeneous data groups, or clusters. | Your firm is in the process of bringing products A and B to market. Modeler - Segmentation/Clustering allows you to:  
  - Regroup your customers into several homogeneous groups.  
  - Understand the behavior of each of these groups with respect to products A and B. |

2.2 Which Sections should you Read?

Depending on your job profile and your needs, you may choose to read the entire guide, or only certain sections. In either case, it is essential that you read the section concerning the Automated Analytics performance indicators. These indicators embody one of the most important concepts the application: they allow evaluation of the quality and robustness of the models generated.

The following table provides some points of reference to facilitate your use of this guide.

<table>
<thead>
<tr>
<th>Your Profile</th>
<th>Best Use of this Guide</th>
</tr>
</thead>
</table>

Classification, Regression, Segmentation and Clustering Scenarios
Welcome to this Guide
<table>
<thead>
<tr>
<th>You want to evaluate Automated Analytics and your time is tightly budgeted</th>
<th>You could restrict yourself to:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Reading the scenario of the feature that interests you (or at least, the summary of that scenario):</td>
<td></td>
</tr>
<tr>
<td>● Application scenario for regression/classification:: Enhance efficiency and master your budget using modeling</td>
<td></td>
</tr>
<tr>
<td>● Application scenario for segmentation/clustering: Customize your communications using data modeling</td>
<td></td>
</tr>
<tr>
<td>2. Going directly to the relevant section:</td>
<td></td>
</tr>
<tr>
<td>● Modeler - Regression/Classification</td>
<td></td>
</tr>
<tr>
<td>● Modeler - Segmentation/Clustering</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>You want to be guided step-by-step through the application</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. You have had only limited hands-on experience in data modeling</td>
</tr>
<tr>
<td>2. Reading all sections of the guide through at least once, in the order in which they are presented.</td>
</tr>
<tr>
<td>Ensure that you have a complete grasp of the essential concepts relating to the use of the application, by reading the chapter on Essential Concepts. These concepts are essential both for the use of the application features and for analysis of the results obtained.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>You have had significant experience in data modeling</th>
<th>You could limit yourself to:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Verifying that you are familiar with the terminology used by the application, by examining the contents of chapter Essential Concepts.</td>
<td></td>
</tr>
<tr>
<td>2. Reading the summary of the scenario of the feature that interests you:</td>
<td></td>
</tr>
<tr>
<td>● Application scenario for regression/classification:: Enhance efficiency and master your budget using modeling</td>
<td></td>
</tr>
<tr>
<td>● Application scenario for segmentation/clustering: Customize your communications using data modeling</td>
<td></td>
</tr>
<tr>
<td>3. Going directly to the relevant section:</td>
<td></td>
</tr>
<tr>
<td>● Modeler - Regression/Classification</td>
<td></td>
</tr>
<tr>
<td>● Modeler - Segmentation/Clustering</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>You have previously taken a training seminar</th>
<th>You can:</th>
</tr>
</thead>
<tbody>
<tr>
<td>You can:</td>
<td></td>
</tr>
<tr>
<td>Follow the application scenarios for a &quot;review&quot; of the features that interest you:</td>
<td></td>
</tr>
<tr>
<td>● Application scenario for regression/classification:: Enhance efficiency and master your budget using modeling</td>
<td></td>
</tr>
<tr>
<td>● Application scenario for segmentation/clustering: Customize your communications using data modeling</td>
<td></td>
</tr>
</tbody>
</table>

Use this document as a reference text, consulting it as required. In this case, the detailed table of contents and the index will be valuable tools, helping you find the information that you seek.
2.3 Files and Documentation Provided with this Guide

Sample data files are supplied with the application. These files allow you to take your first steps using various features of the application, and evaluate them. During installation, the sample files are saved in the folder: C:\Program Files\SAP BusinessObjects Predictive Analytics\Desktop \X.Y\Automated\Samples\Census\.

The following table describes those files.

<table>
<thead>
<tr>
<th>File Name</th>
<th>Description</th>
<th>When is it Used?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Census01.csv</td>
<td>Data file</td>
<td>This file is used for both application scenarios used in this manual.</td>
</tr>
<tr>
<td>desc_census01.csv</td>
<td>Description file for the Census01.csv file</td>
<td>This file is used for both application scenarios used in this manual.</td>
</tr>
</tbody>
</table>

To obtain a detailed description of the Census01.csv file, see the related topic.

Complete documentation is included with the application. This documentation covers:

- The operational use of the application features
- The architecture and integration of the application API
- The Java graphical user interface

Each screen in the modeling assistant is accompanied by contextual help that describes the options presented to you, and the concepts required for their application. To display the contextual help, press F1.

To access a searchable version of all the help for the application, select Help Open Full Searchable Help.

Related Information

Introduction to Sample Files [page 163]
3 Overview of Automated Analytics

Automated Analytics provides a data mining solution for modeling your data as easily and rapidly as possible, while maintaining relevant and readily interpretable results. You will transform your data into knowledge in order to make timely strategic and operational decisions.

The application places the latest data mining techniques within reach of any non-expert user. You can access many data source formats to rapidly generate explanatory and predictive models, as well as descriptive models in a semi-automated manner.

The application allows you to concentrate on high value-added activities such as analysis of the results of data modeling, and decision-making.

3.1 Automated Analytics Architecture

The figure below illustrates the general architecture of Automated Analytics.

Three types of interfaces allow you to use the application:
- Graphical user interface,
- Command interpreter,
- API (Application Programming Interface) controls.

The graphical user interface is aimed primarily at non-expert users. It allows you to model your data very easily. In addition, it provides plotted output to facilitate viewing and interpretation of the results of modeling.

The KxShell command interpreter allows you to use the application by typing commands or executing scripts containing several commands. The command interpreter is an example of development based on the C++ API. Like any other API, it may be used to integrate the application with other applications or program packages.

The Control API is aimed primarily at developers, or users with programming experience. This API is used to access the complete range of functionalities and the most fine-grained parameterization of the application’s
features. In addition, it allows customized integration of the application’s features with other applications or program packages.

Three APIs are provided with the application:

- A COM/DCOM API, usable on Microsoft platforms,
- A CORBA API, usable on all client/server platforms,
- A C++ API, usable on all standalone platforms.

### 3.2 Automated Analytics Operational Phases

The operation of the Automated Analytics can be subdivided into four phases:

- Data access
- Data manipulation and preparation
- Data modeling
- Model presentation and deployment

#### Data Access Phase

The application accepts many types of data sources:

- Flat files such as .csv files, files of text tables and other files of type text.
- ODBC-compatible sources such as Oracle, SQL Server or IBM DB2 databases.

In addition, the C Data Access Application Programming Interface allows you to connect proprietary format sources, such as industrial sensor streams.

In most cases, and particularly if you are using the application features via a graphical interface, you never have to concern yourself with the data access process. Data access is accomplished in a semi-transparent manner: from the graphical user interface, you need only select the data source format to be used (flat files or ODBC-compatible data sources) and specify the location of the data file. The C Data Access Application Programming Interface is helpful to developers who want to write access code for proprietary format databases.

The application allows reading SAS data and writing the scores obtained with an application model into a SAS table.

The following formats are currently supported:

- SAS files version 6 under Windows and UNIX
- SAS 7/8 under Windows and UNIX
- SAS Transport Files

You can directly access a SAS table with the application interface by selecting the format of the file to analyze. Once you have built your model with the application, you can generate a SAS table containing the model application results (for example, scores, probability, cluster number, predicted value). The application interface allows you to select the output format. The generated SAS table is automatically integrated in SAS information system.
Data Manipulation and Preparation Phase

The Sequence Coding and Event Logging features of Data Manager are data manipulation and preparation features. They are used to encode data in a robust and semi-automatic manner, making them available for use by all analytical features of the application. The use of these features is transparent: all data processing is performed automatically.

Event Logging (formerly known as KEL) aggregates events into periods of time. It allows integrating transactional data with demographic customer data. It is used in cases when the raw data contains static information such as age, gender or profession of an individual, and dynamic variables, such as spending patterns or credit card transactions. Data is automatically aggregated within user defined periods without programming SQL or changing database schema. Event Logging combines and compresses this data to make it available to other features of the application.

Sequence Coding (formerly known as KSC) aggregates events into a series of transitions. For example a customer click-stream from a Web site can be transformed into a series of data for each session. Each column represents a specific transition from one page to another. Similar to Event Logging these new columns of data can be added to existing customer data and are made available to other application features for further processing.

Modeler - Data Encoding (formerly known as K2C) automatically prepares and transforms data into a format suitable for use in the application. Modeler - Data Encoding translates nominal and ordinal variables, automatically fills in missing values and detects out of range data. In addition, this feature contributes significantly to the robustness of the models generated by the application engine, by providing a robust data encoding.

Data Modeling Phase

Thanks to the statistical techniques and information technologies upon which the Regression/Classification, Segmentation/Clustering and Time Series features were built, these features require only a short modeling time to generate relevant and robust analytical models of your data.

Modeler - Regression/Classification (formerly known as K2R) generates explanatory and predictive models. The models generated by Classification/Regression explain and predict a phenomenon, or business question, by a function of the analyzed dataset, the explanatory variables. The models generated are calculated using a regression and classification algorithm. This polynomial regression is a proprietary algorithm using Vapnik’s SRM (Structural Risk Minimization) principle to calculate the parameters.

Modeler - Segmentation/Clustering (formerly known as K2S) generates descriptive models, which means a function to regroup cases in a dataset into a number of clusters with similar behavior toward a business question.

Time Series (formerly known as KTS) lets you build predictive models from data representing time series. Thanks to time series models, you can:

- Identify and understand the phenomenon represented by your time series.
- Forecast the evolution of time series in the short and medium term, that is, predict their future values.
Model Presentation and Deployment Phase

Once the models have been generated, model performance indicators, plots and modeling reports in HTML format facilitate viewing and interpretation of the data modeling results.

Once the models have been validated, you can apply them to:

- One or more specific observations taken from your database (Simulation mode).
- A new, complete dataset or application dataset (Batch mode).

To facilitate deployment and integration of the models, the code corresponding to each model can also be generated in the programming language using the code generation feature.

The code generation feature (formerly known as Scorer or KMX) generates code in the following languages: C, XML, AWK, HTML, SQL, PMML2, SAS, or JAVA corresponding to a model generated by the application.

In this form, the model may be integrated into any application that supports the aforementioned languages.

The generated codes allow the application models to be integrated within any given application or software package, or to be applied directly to the data without requiring the Automated Analytics environment.

**i Note**

Code generation is only available for models using the following features: Modeler - Data Encoding, Modeler - Regression/Classification, Modeler - Segmentation/Clustering.

3.3 Methodological Prerequisites

Before modeling your data using Automated Analytics, you should:

- State a business issue that you want to solve,
- Possess a dataset representing this issue in the form of a set of observations.

What is your Business Issue?

The application is designed to allow supervised data analysis. The term supervised means that the data analysis does not occur completely independently, but always as a function of a particular issue: your business issue.

Consider the database that contains information about your customers. An analysis that groups your customers into homogeneous groups independently of your input is of little interest. On the other hand, an analysis that groups them as a function of a variable such as “mean business revenues earned from this customer each year” offers significant interest. You would learn the characteristic profiles of the customers that bring you the most money. Then, you can develop strategies to better influence your customers according to their characteristic profiles.

Before you begin, you need to identify and formulate your business issue.
Is your Data Usable?

Once your business issue has been identified and formulated, you need to have data on hand that will permit an answer to be found. In order for your data to be usable by the application, the following five conditions must be met:

- You must have a sufficiently large volume of data to be able to build a valid model, that is, in order for the model to be both relevant and robust. An analytical model that is generated from a dataset of 50 lines may have low generalization capacity, and contain low informative value. We can advise you on the issues of data volume.
- Your dataset must contain a target variable, that will allow you to express your business issue within the application.
- The target variable must be known for each observation of the training dataset. To express this another way, no target variable values may be missing over the range of the entire training dataset.
- The data source format must be supported by the application.
- Your data must be presented in the form of a single table of data, except in instances where you are using Event Logging or Sequence Coding features of Data Manager.
4 Essential Concepts

4.1 Operation of Automated Analytics: Overview

Automated Analytics allows you to perform supervised data mining, that is, to transform your data into knowledge, then into action, as a function of a domain-specific business issue.

The application supports various formats of source data (flat files, ODBC-compatible sources). In order to be usable by the application features, the dataset to be analyzed must be presented in the form of a single table of data, except in instances where you are using the Event Logging or Sequence Coding features of Data Manager.

To use the application’s features, you must have a training dataset available that contains the target variable with all its values defined. Then, you can apply the model generated using the training dataset to one or more application datasets.

The training dataset is cut into three data sub-sets for estimation, validation and testing, using a cutting strategy.

The different types of variables: continuous, ordinal and nominal are next encoded by the data encoding feature of Modeler, or by the Event Logging and Sequence Coding features in the case of dynamic data. Before generating the model, you must:

- Describe the data. A utility integrated with the application allows you to generate a description of the dataset to be analyzed, automatically. You need only validate that description, verifying that the type and storage format of each variable were identified correctly.
- Define the role of variables contained in the dataset to be analyzed. You may select one or more variables as target variables. These are the variables that corresponds with your business issue. The other variables of the table of data are considered to be explanatory variables: they allow calculation of the value of the target variable in a given context. They may also be used as weight variables.

For more information about the role of each feature of the application, see the description of the operational phases.

You can then generate models, capable of either explaining and predicting a phenomenon, or describing a dataset, in both cases, as a function of the previously defined target variable. This phase is called the training phase.

Once the models have been generated, you can view and interpret their relevance and robustness using:

- Performance indicators: the predictive power, which is the quality indicator, and the prediction confidence, which is the robustness indicator.
- A variety of plots, including the profit curve plot.

Related Information

Automated Analytics Operational Phases [page 13]
Performance Indicators [page 38]
4.2 Data Sources Supported

Automated Analytics supports the following data sources:

- Text files (also called flat files) in which the data are separated by a delimiter, such as commas in .csv (Comma Separated Value) files.

  **Restriction**
  
  When accessing data in .csv files, Automated Analytics only supports CR + LF (common on Microsoft Windows) or LF (common on UNIX) for line breaks.

- Database management systems that can be accessed using ODBC.

  **Note**
  
  For the list of supported ODBC-compatible sources, see the SAP Product Availability Matrix (PAM) at http://service.sap.com/sap/support/pam.
  
  For more information about using SAP HANA, see the related information below.
  
  To configure Automated Analytics modeling tools to access data in your database management system, refer to the guide Connecting your Database Management System on Windows or Connecting your Database Management System on UNIX.

- SAS files

Related Information

SAP HANA as a Data Source [page 18]

4.2.1 SAP HANA as a Data Source

You can use SAP HANA databases as data sources for all types of modeling analyses in Modeler:

- Classification/Regression
- Clustering
- Time Series
- Association Rules
- Social
- Recommendation

| SAP HANA tables or SQL views | found in the Catalog node of the SAP HANA database |
All types of SAP HANA views

An SAP HANA view is a predefined virtual grouping of table columns that enables data access for a particular business requirement. Views are specific to the type of tables that are included, and to the type of calculations that are applied to columns. For example, an analytic view is built on a fact table and associated attribute views. A calculation view executes a function on columns when the view is accessed.

**Restriction**
- Analytic and calculation views that use the variable mapping feature (available starting with SAP HANA SPS 09) are not supported.
- You cannot edit data in SAP HANA views using Automated Analytics.
- SAP HANA views are not supported in Data Manager.

Smart Data Access virtual tables

Thanks to Smart Data Access, you can expose data from remote sources tables as virtual tables and combine them with HANA regular tables. This allows you to access data sources that are not natively supported by the application, or to combine data from multiple heterogeneous sources.

**Caution**
To use virtual tables as input data sets for training or applying a model or as output data sets for applying a model, you need to check that the following conditions are met:
- The in-database application mode is not used.
- The destination table for storing the predicted values exists in the remote source before applying the model.
- The structure of the remote table, that is the column names and types, must match exactly what is expected with respect to the generation options; if this is not the case an error will occur.

**Caution**
In Data Manager, use virtual tables with caution as the generated queries can be complex. Smart Data Access may not be able to delegate much of the processing to the underlying source depending on the source capabilities. This can impact performance.
Prerequisites

You must know the ODBC source name and the connection information for your SAP HANA database. For more information, contact your SAP HANA administrator.

In addition to having the authorizations required for querying the SAP HANA view, you need to be granted the SELECT privilege on the _SYS_BI schema, which contains metadata on views. Please refer to SAP HANA guides for detailed information on security aspects.

4.2.2 Managing Performance when Using Databases

Before requesting data stored in a Teradata(1), Oracle(2) or SQLServer 2005 database, the application uses a feature, called the Explain mode, which categorizes the performances of SQL queries in several classes defined by the user. In order to be as fast and as light as possible, this categorization is done without actually executing the full SQL query.

Note

- (1) For all versions of Teradata.
- (2) For all versions above and including Oracle 10.

The objective is to allow estimating the workload of the SQL query before executing it and then deciding -- possibly thanks to an IT Corporate Policy -- if the SQL query can actually be used.

For example, an IT Corporate Policy may favor interactivity and then define 3 classes of SQL queries, each with its maximum time:

- **Immediate**: duration < 1 s. The query is accepted and executed immediately.
- **Batched**: 1 s <= duration < 2 s. The query is accepted but will be executed on next idle time.
- **Rejected**: 2 s <= duration. The query will never be executed.

The number, names and limits of classes are defined by the user in order for these values to match the current DBMS configuration and DBMS usage policy.

The Explain Mode has been Configured

If the Explain mode has been configured by your DBMS administrator, there are two possible outcomes to a query:

- **the query is accepted and executed**: this is completely transparent. The application accesses the data without further input from the user.
- **the query needs to be validated before being executed**: a pop-up window opens displaying a message configured by the DBMS administrator. A query that needs validation can be categorized in two ways:

  medium-sized or huge.
If the query is categorized as medium-sized, you will probably have to check with you administrator which action to take:

- If the administrator authorizes the query, click **Continue**. The pop-up window closes and the requested action is carried out.
- If the administrator does not authorize the query, click **Stop Query**, the pop-up window closes, but no action is executed.

If the query is categorized as huge, it means that the query will take too much time and resources. In that case, the behavior of the **Continue** button depends on the configuration set by the DBMS Administrator (for example, it can automatically refuse queries that are considered too heavy). In any case, you should check with them to know the line of action to follow.

### The Explain Mode has not been Configured

If your DBMS Administrator has not configured the Explain mode, the following pop-up opens when you try to access the data:

You need to contact your Administrator who will tell you which action to take and configure the Explain mode.

If the Administrator validates the execution of the query, you may want all queries with the same duration to be executed without validation. In that case, check the box **Do not request validation anymore for similar requests**. The validation message will then only appear for larger queries. This configuration will only be used for the current session, when closing the application, it will be lost. For a permanent configuration, see your DBMS Administrator.

### 4.3 Training and Application Dataset

To use the application’s features, you must have a training dataset available that contains the target variable with all its values defined. Then, you can apply the model generated using the training dataset to one or more application datasets.

#### Training Dataset

A training dataset is a dataset used for generating a model. In this set, the values of the target variable – or variable corresponding to your business issue – are known. By analyzing the training dataset, the application
features will generate a model that allows explanation of the target variable, based on the explanatory variables.

To allow validation of the model generated, the training dataset is cut into three sub-sets using a cutting strategy.

The training dataset may correspond to either a complete population section of your database or a sample extracted from this population. The choice depends on the type of study to be performed, the tools used and the budget allocated to the study.

Application Dataset

An application dataset is a dataset to which you apply a model. This dataset contains an unknown target variable for which you want to know the value.

The model applied to the application dataset must have been previously generated from a training dataset. The application dataset must contain exactly the same information structure as the corresponding training dataset, that is:

- The same number of variables,
- The same types of variables,
- The same order of presentation of these variables.

Caution

The application dataset must contain a target variable that corresponds to that of the training dataset. This point is true for all instances, even if the values of this target variable are empty. When these values are defined, they may serve to detect any possible deviant observations (outliers).

Related Information

Cutting Strategies [page 22]
Target Variable [page 32]

4.4 Cutting Strategies

A cutting strategy is a technique that allows decomposition of a training dataset into three distinct sub-sets:

- An estimation sub-set
- A validation sub-set
- A test sub-set

This cutting allows for cross-validation of the models generated.

The following table defines the roles of the three data sub-sets obtained using cutting strategies.
### The Data-Set

<table>
<thead>
<tr>
<th>The data-set</th>
<th>Is used to...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimation</td>
<td>Generate different models. The models generated at this stage are hypothetical.</td>
</tr>
<tr>
<td>Validation</td>
<td>Select the best model among those generated using the estimation sub-set, which represents the best compromise between perfect quality and perfect robustness.</td>
</tr>
<tr>
<td>Test</td>
<td>Verify the performance of the selected model on a new dataset.</td>
</tr>
</tbody>
</table>

To understand the role of cutting strategies in the model generation process, see the related topic on generating a model.

To generate your models, there are two types of cutting strategies that you may use:

- The customized cutting strategy
- The automatic cutting strategies

### Related Information

- Generating a Model [page 35]
- Customized Cutting Strategy [page 23]
- Automatic Cutting Strategies [page 23]

### 4.4.1 Customized Cutting Strategy

The customized cutting strategy allows you to define your own data sub-sets. To use this strategy, you must have prepared (before opening the application features) three sub-sets: the estimation, validation and test sub-sets.

Before opening the application, cut your initial data file into three files of the size of your choice. For example:

- The first file may contain the first 1,500 observations, or lines, of your initial data file,
- The second file, observations 1,501 to 3,000,
- The third file, observations 3,001 to 5,000.

⚠ **Caution**

The customized cutting strategy is risky in the instance of an initial data file in which the data have been sorted. In this case, the first lines will not be representative of the overall set of data contained in the first file. To avoid this type of bias, do not forget to mix up your data prior to analysis.

### 4.4.2 Automatic Cutting Strategies

With the exception of the customized cutting strategy, cutting strategies are automatic. Automatic cutting strategies operate upon a single data file, which constitutes your initial dataset.
Automatic cutting strategies always cut the initial dataset into the same proportions. The following table details the proportions attributed to each dataset depending on the presence of a test dataset.

<table>
<thead>
<tr>
<th>Automatic Cutting Strategies with Test</th>
<th>Automatic Cutting Strategies without Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>● 3/5 of the data are used in the estimation sub-set,</td>
<td>● 3/4 of the data are used in the estimation sub-set,</td>
</tr>
<tr>
<td>● 1/5 of the data are used in the validation sub-set,</td>
<td>● 1/4 of the data are used in the validation sub-set.</td>
</tr>
<tr>
<td>● 1/5 of the data are used in the test sub-set.</td>
<td></td>
</tr>
</tbody>
</table>

There are seven types of automatic cutting strategies available within the application. The random cutting strategies distribute the data of the initial dataset in a random manner between the three sub-sets: estimation, validation and test.

### Random with Test at the End

The Random with test at the end cutting strategy distributes:

- 4/5 of the initial dataset in a random manner in the two sub-sets estimation and validation, 3/5 being distributed in the estimation data sub-set and 1/5 in the validation data sub-set.
- The final 1/5 of the initial dataset is sent directly into the test sub-set.

This is a useful strategy in cases where:

- Your database corresponds to a well-defined evolution because of the way it was built, which may mean, for example, that the data is in chronological order,
- You may wish to take this order into account when generating your model.

For example, imagine that:

- New customers are added every month to your database,
- You know that the datasets to which you apply the model will, once generated, have a better chance of resembling the most recent section of your database, that is, the section that contains the most recent customers entered.

Using the Random with test at the end cutting strategy, you decide to test the model generated on that section of your database that is most likely to resemble the state of your future application datasets.
Random Without Test (Default Strategy)

The Random without test strategy is the cutting strategy suggested as the default setting. It distributes the whole initial dataset in a random manner to the two sub-sets of estimation and validation:

- 3/4 of the initial dataset are distributed to the estimation sub-set.
- 1/4 to the initial dataset are distributed to the validation sub-set.

As no test sub-set is used, all the data from your training dataset can be used for sub-sets of estimation and validation. This can lead to a model with a better quality and robustness.

Periodic

The Periodic cutting strategy is implemented by following this distribution cycle:

1. Three lines of the initial dataset are distributed to the estimation sub-set.
2. One line is distributed to the validation sub-set.
3. One line is distributed to the test sub-set.
4. Distribution begins again at step 1.

Periodic with Test at the End

The Periodic with test at the end strategy distributes:

- 4/5 of the initial dataset in a periodic manner to the two sub-sets of estimation and validation, 3/5 being distributed in the estimation data sub-set and 1/5 in the validation data sub-set 3/5 being distributed.
- The final 1/5 of the initial dataset is sent as a block of data to the test sub-set.

In other words, this strategy follows this distribution cycle:

1. Three lines of the first 4/5 of the initial dataset are distributed to the estimation sub-set.
2. One line of the first 4/5 of the initial dataset is distributed to the validation sub-set.
3. If the entire 4/5 of the initial dataset is not yet distributed, distribution operations begin again at step 1. Otherwise, if the entire 4/5 of the initial dataset has been distributed, distribution operations go to step 4.
4. The final 1/5 of the initial dataset is sent as a block of data to the test sub-set.

Periodic Without Test

The Periodic without test strategy distributes the whole initial dataset in a periodic manner to the two sub-sets of estimation and validation:

- 3/4 of the initial dataset are distributed to the estimation sub-set.
- 1/4 to the initial dataset are distributed to the validation sub-set.
In other words, this cutting strategy is implemented by following this distribution cycle:

1. Three lines of the initial dataset are distributed to the estimation sub-set.
2. One line is distributed to the validation sub-set.
3. Distribution begins again at step 1.

As no test sub-set is used, all the data from your training dataset can be used for sub-sets of estimation and validation. This can lead to a model with a better quality and robustness.

**Sequential**

The Sequential strategy cuts the initial dataset into three blocks, corresponding to the usual cutting proportions:

- The lines corresponding to the first 3/5 of the initial dataset are distributed as a block to the estimation dataset.
- The lines corresponding to the next 1/5 of the initial dataset are distributed as a block to the validation dataset.
- The lines corresponding to the final 1/5 of the initial dataset are distributed as a block to the test dataset.

**Sequential without Test**

The Sequential without test strategy cuts the initial dataset into two blocks:

- The lines corresponding to the first 3/4 of the initial dataset are distributed as a block to the estimation dataset.
- The lines corresponding to the next 1/4 of the initial dataset are distributed as a block to the validation dataset.

As no test sub-set is used, all the data from your training dataset can be used for sub-sets of estimation and validation. This can lead to a model with a better quality and robustness.

### 4.5 Table of Data

A table of data is a dataset presented in the form of a two-dimensional table.

In this table:

- Each row represents an observation to be processed, such as "American individual" in the sample file Census01.csv.
- Each column represents a variable that describes observations, such as the "age" or the "gender" of individual Americans.
- Each cell, the intersection of a column and a row, represents the value of the variable in the column, for the observation in that row.

The following table is an example of a table of data.
Observations | Variable 1 | Variable 2 | Variable 3
---|---|---|---
Observation a | Value a1 | Value a2 | Value a3
Observation b | Value b1 | Value b2 | Value b3
... | ... | ... | ...
Observation n | Value n1 | Value n2 | Value n3

### Synonyms of "Observation" and "Variable"

Depending upon your profile and your area of expertise, you may be more familiar with other terms that refer to observations (in rows) and variables (in columns) when using tables of data.

The following table presents such terms, or synonyms.

<table>
<thead>
<tr>
<th>Terms equivalent to the term &quot;Observation&quot;</th>
<th>Terms equivalent to the term &quot;Variable&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Row</td>
<td>Column</td>
</tr>
<tr>
<td>Record</td>
<td>Attribute</td>
</tr>
<tr>
<td>Table</td>
<td>Field</td>
</tr>
<tr>
<td>Event</td>
<td>Property</td>
</tr>
<tr>
<td>Instance</td>
<td>-</td>
</tr>
<tr>
<td>Example</td>
<td>-</td>
</tr>
</tbody>
</table>

### Data Formats

Whatever the data source used, the following two constraints must be accommodated:

- The data must be represented in the form of a single table, except in instances where you are using the Event Logging or Sequence Coding features.
- The target variable must be defined for each observation in the table. In the sample file Census01.csv, the variable "class" has been defined for each individual.

**Note**

For the list of supported ODBC-compatible sources, see the SAP Product Availability Matrix (PAM) at [http://service.sap.com/sap/support/pam](http://service.sap.com/sap/support/pam).
4.6 Variables

A variable corresponds to an attribute which describes the observations stored in your database. For example, in a database containing information about your customers, the "name" and "address" of those customers are examples of variables. In Automated Analytics, a variable is defined by three aspects:

- The type of variable:

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuous</td>
<td>Variables whose values are numerical, continuous, and sortable. They can be used to calculate measures (such as mean or variance).</td>
</tr>
<tr>
<td>Ordinal</td>
<td>Variables with discrete values that are sortable.</td>
</tr>
<tr>
<td>Nominal</td>
<td>Variables with discrete values that are not sortable.</td>
</tr>
<tr>
<td>Textual</td>
<td>A type of nominal variable containing phrases, sentences or complete texts. Textual variables are used for text analyses.</td>
</tr>
</tbody>
</table>

- The storage format of the variable (for example, date, number, or string)
- The role of the variable: Target, Explanatory, or Weight

Related Information

Continuous Variables [page 28]
Ordinal Variables [page 29]
Nominal Variables [page 29]
Variable Storage Formats [page 30]
Target Variable [page 32]
Explanatory Variable [page 33]
Weight Variable [page 34]

4.6.1 Continuous Variables

Continuous variables are variables whose values are numerical, continuous and sortable. Arithmetic operations may be performed on these values, such as determination of their sum or their mean.

During modeling, a continuous variable may be grouped into significant discrete bins.

Example

The variable "salary" is a numerical variable, but in addition, is also a continuous variable. It may, for instance, take on the following values: "$1,050", "$1,700", or "$1,750". The mean of these values may be calculated.
4.6.2 Ordinal Variables

Ordinal variables are variables with discrete values, that is, they belong to categories, and they are sortable. Ordinal variables may be:

- Numerical, meaning that its values are numbers. They are therefore ordered according to the natural number system (0, 1, 2, and so on).
- Textual, meaning that its values are character strings. They are therefore ordered according to alphabetic conventions.

**Example**

The variable "school grade" is an ordinal variable. Its values actually belong to definite categories and can be sorted. This variable can be:

- numerical, if its values range between "0" and "20",
- textual, if its values are A, B, C, D, E et F.

**Caution**

A variable "assessment" which can have the values "good", "average", and "bad", cannot be directly treated as an ordinal variable by the application. The values would be sorted in alphabetical order ("average","bad", "good") and not according to their meaning. When a nominal variable order is important, the variable must be encoded, in letters or numbers, before it can be used by the application.

4.6.3 Nominal Variables

Nominal variables are variables whose values are discrete, that is, belong to categories, and are not sortable. Nominal variables may be:

- Numerical, meaning that its values are numbers.
- Textual, meaning that its values are character strings.

**Caution**

Binary variables are considered nominal variables.

During modeling, the values of the categorical variables are regrouped into homogeneous categories. These categories are then ordered as a function of their relative contribution with respect to the values of the target variable.

**Example**

The variable "zip code" is a nominal variable. The set of values that this variable may assume ("10111", "20500", "90210", for example) are clearly distinct, non-ranked categories, although they happen to be represented by numbers.
The variable “eye color” is a nominal variable. The set of values that this variable may assume (“blue”, “brown”, “black”, for example) are clearly distinct, non-ordered categories, and are represented by character strings.

### 4.6.4 Variable Storage Formats

To describe the data, the application uses four types of storage formats:

- **date**
- **datetime**
- **number**
- **integer**
- **string**

The following table describes these storage formats.

<table>
<thead>
<tr>
<th>The storage format...</th>
<th>Is used to describe variables when their values correspond to...</th>
<th>For instance...</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>date</strong></td>
<td>Dates expressed in the following formats:</td>
<td>&quot;2001-11-30&quot;</td>
</tr>
<tr>
<td></td>
<td>- YYYY-MM-DD</td>
<td>&quot;1999/04/28&quot;</td>
</tr>
<tr>
<td></td>
<td>- YYYY/MM/DD</td>
<td></td>
</tr>
<tr>
<td><strong>datetime</strong></td>
<td>Dates and times expressed in the following formats:</td>
<td>&quot;2001-11-30 14:08:17&quot;</td>
</tr>
<tr>
<td><strong>number</strong></td>
<td>Figures, or numerical values on which operations may be performed</td>
<td>The variable &quot;salary&quot;, in US dollars: &quot;1000.00&quot;, &quot;1593&quot; and &quot;2000.54&quot;</td>
</tr>
<tr>
<td><strong>integer</strong></td>
<td>Figures, or numerical integer values on which operations may be performed</td>
<td>The variable &quot;age&quot;, in years: &quot;21&quot;, &quot;34&quot; and &quot;99&quot;</td>
</tr>
<tr>
<td><strong>string</strong></td>
<td>Alphanumeric character strings</td>
<td>The variable “family name”: “Lake”, &quot;Martin&quot; and &quot;Miller&quot;</td>
</tr>
<tr>
<td></td>
<td>- The variable “occupation”: “professor”, “engineer” and “translator”</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- The variable “telephone”: “800 555 1234” and “800 555 4321”</td>
<td></td>
</tr>
</tbody>
</table>

**Note**

A variable that has numbers for values is not forced to be described using the number storage format. For instance, the variables “telephone” and “zip code” may instead be described using the string storage format, because no arithmetic operations that make any sense can be performed on these values. Similarly, a variable that will be used as an observation identification code in a table, and does not comply with supported number formats may be described using the string storage format.
Caution

For number storage formats, the decimal separator used must be a decimal point, and not a comma. So, the value “6.5” may be processed, while “6,5” will not be processed.

Date and Datetime Variables: Automatically Generated Variables

When your dataset contains date or datetime variables, the application automatically extracts date information as shown in the following tables.

For date or datetime variables:

<table>
<thead>
<tr>
<th>Temporal Information</th>
<th>Represents...</th>
<th>Generated Variable Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Day of week</td>
<td>the day of week according to the ISO disposition (Monday=0 and Sunday=6)</td>
<td><code>&lt;OriginalVariableName&gt;_DoW</code></td>
</tr>
<tr>
<td>Day of month</td>
<td>the day of month (1 to 31)</td>
<td><code>&lt;OriginalVariableName&gt;_DoM</code></td>
</tr>
<tr>
<td>Day of year</td>
<td>the day of the current year (1 to 366)</td>
<td><code>&lt;OriginalVariableName&gt;_DoY</code></td>
</tr>
<tr>
<td>Month of quarter</td>
<td>the month of the quarter (January, April, July and October = 1, February, May, August and November = 2, March, June, September and December = 3)</td>
<td><code>&lt;OriginalVariableName&gt;_MoQ</code></td>
</tr>
<tr>
<td>Month of year</td>
<td>the month (1 to 12)</td>
<td><code>&lt;OriginalVariableName&gt;_M</code></td>
</tr>
<tr>
<td>Year</td>
<td>the year</td>
<td><code>&lt;OriginalVariableName&gt;_Y</code></td>
</tr>
<tr>
<td>Quarter</td>
<td>the quarter of the year (January to March = 1, April to June = 2, July to September = 3, October to December = 4)</td>
<td><code>&lt;OriginalVariableName&gt;_Q</code></td>
</tr>
</tbody>
</table>

For datetime variables:

<table>
<thead>
<tr>
<th>Temporal Information</th>
<th>Represents...</th>
<th>Generated Variable Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hour</td>
<td>the hour</td>
<td><code>&lt;OriginalVariableName&gt;_H</code></td>
</tr>
<tr>
<td>Minute</td>
<td>the minute</td>
<td><code>&lt;OriginalVariableName&gt;_Mi</code></td>
</tr>
<tr>
<td>Second</td>
<td>the second</td>
<td><code>&lt;OriginalVariableName&gt;_S</code></td>
</tr>
<tr>
<td>µ seconds</td>
<td>the micro-second</td>
<td><code>&lt;OriginalVariableName&gt;_mu</code></td>
</tr>
</tbody>
</table>

The generated variables will appear in the model debriefing panels listing variables, such as the Contributions by Variable, the Category Significance, the Statistical Reports, as well as in the automatic variable selection feature.
4.6.5 Target Variable

A target variable is the variable that you seek to explain, or for which you want to predict the values in an application dataset. It corresponds to your domain-specific business issue.

When the target variable is a binary variable, the application considers that the target value, or target category, of this variable (that is, the value that is the object of the analysis) to be the least frequently occurring value in the training dataset. Imagine that a training dataset containing the customer information of a company contains the target variable "responded to my mailing". This target variable may take the values "Yes" or "No". If the value "Yes" is the least frequent value (for instance, if 40% of referenced customers responded to the mailing), the application considers that value to be the target category of the target variable.

Synonyms

Depending upon your profile and your area of expertise, you may be more familiar with one of the following terms to refer to target variables:

- Variables to be explained,
- Dependent variables,
- Output variables.

These terms are synonyms.

Constraints Governing Use

The following constraints govern the use of a target variable:

- Within a training data set, all target variable values must be known.
- Only binary or continuous variables may be used as target variables.

Example

Your company is marketing two products A and B.

You have a database which contains references to:

- 1,500 of your customers. You know which product, A or B, each customer has purchased.
- 10,000 prospects. You want to know which product each customer is likely to purchase.

The variable "product purchased" is your target variable: it corresponds to your business issue. It is:

- Known for all values of the training dataset (in our example, the customers),
- Not known for the values of the application dataset (in our example, the prospects).

The application features allow you to model that target variable, and thus predict which product each of your prospects is likely to purchase.
The following table represents your database.

<table>
<thead>
<tr>
<th>Name</th>
<th>Age</th>
<th>Residence</th>
<th>Socio-Occupational Category</th>
<th>Product Purchased</th>
</tr>
</thead>
<tbody>
<tr>
<td>Charles</td>
<td>34</td>
<td>New Orleans</td>
<td>Manager/Administrator</td>
<td>Product A</td>
</tr>
<tr>
<td>John</td>
<td>37</td>
<td>Washington</td>
<td>Manager/Administrator</td>
<td>Product A</td>
</tr>
<tr>
<td>Marlène</td>
<td>31</td>
<td>Boston</td>
<td>Civil servant</td>
<td>Product B</td>
</tr>
<tr>
<td>Prospect 1</td>
<td>34</td>
<td>Oakland</td>
<td>Manager/Administrator</td>
<td>?</td>
</tr>
<tr>
<td>Prospect 2</td>
<td>24</td>
<td>Washington</td>
<td>Civil servant</td>
<td>?</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Prospect n</td>
<td>35</td>
<td>Sacramento</td>
<td>Skilled tradesman</td>
<td>?</td>
</tr>
</tbody>
</table>

4.6.6 Explanatory Variable

An explanatory variable is a variable that describes your data and which serves to explain a target variable.

**Synonyms**

Depending upon your profile and your area of expertise, you may be more familiar with one of the following terms to refer to explanatory variables:
- Causal variables,
- Independent variables,
- Input variables.

These terms are synonyms.

**Example**

Your company is marketing two products A and B.

You have a database which contains references to:
- 1,500 of your customers. You know which product, A or B, each customer has purchased.
- 10,000 prospects. You want to know which product each prospect is likely to purchase.

The variables "name", "age", "address" and "socio-occupational class" are your explanatory variables: they allow you to generate a model capable of explaining and predicting the value of the target variable "product purchased".

The following table represents your database.
Name | Age | Address | Socio-Occupational Class | Product Purchased
--- | --- | --- | --- | ---
Charles | 34 | New Orleans | Manager/Administrator | Product A
John | 37 | Washington | Manager/Administrator | Product A
Marlene | 31 | Boston | Civil servant | Product B
Prospect 1 | 34 | Oakland | Manager/Administrator | ?
Prospect 2 | 24 | Washington | Civil servant | ?
... | ... | ... | ... | ...
Prospect n | 35 | Sacramento | Skilled tradesman | ?

### 4.6.7 Weight Variable

A weight variable allows one to assign a relative weight to each of the observations it describes, and actively orient the training process. To declare a variable a weight variable results in creating a number of copies of each of the dataset observations, proportional to the value they possess for that variable.

Specifying a weight variable can be used:

- either to assign a higher weight to a single line
- or to do stratified sampling.

The effect of the weight can be considered as the following: a line with a weight of two in the training dataset is exactly equivalent to having two identical lines with a weight of one.

### Constraints Governing Use

Only positive continuous variables may be used as weight variables.

### Example

Imagine a dataset in which the observations correspond to individual Americans. These observations are described by the variable "age", among others. Defining the variable "age" as a weight variable means that for generation of the model, older individuals will be weighted more heavily than younger individuals.
4.7  Models

The term model carries many different meanings depending on its field of application. In data mining, a model describes and explains the relationships that exist between input data (explanatory variables) and output data (one or more target variables). It allows one to predict and explain phenomena, or to describe them.


A model that has satisfactory performance is one that possesses both:

- High explanatory power, that is, sufficient capacity to explain the target variable. This explanatory power is indicated by the predictive power of the model.
- High robustness, that is, sufficient capacity to repeat the same performance on new datasets containing observations of a similar nature to the training dataset. This explanatory power is indicated by the prediction confidence.

In data mining, there are two types of models:

- Predictive and explanatory models, which allow one to predict and explain phenomena,
- Descriptive models, which allow one to describe datasets.

With the application, you can generate models that are both highly descriptive and highly predictive.

4.7.1 Generating a Model

The analytics are generated during a phase called the training phase, using a training dataset.

Depending on the situation, the training dataset may be cut into three sub-sets:

- An Estimation sub-set.
- A Validation sub-set.
- A Test sub-set.

A cutting strategy determines the way in which the data of the training dataset are distributed across the sub-sets.

The Estimation and Validation sets are used for actual training, and the Test set (sometimes referred to as the "hold-out sample") is used to ensure that the predicted performance is correct.

**Note**

When using the application, the data sub-sets are virtual. The file corresponding to the initial dataset remains intact at all times.

The figure below illustrates the model generation process, known as the training phase.
4.7.2 Representation of a Model

A model may be represented in many different ways, including:

- a decision tree,
- a neural network,
- a mathematical function.

In the application, models are represented in the form of mathematical functions, specifically, polynomials.

Description of the Polynomial

A polynomial may be of degree 1, 2, 3 or greater. By defining the polynomial degree, you are defining the degree of complexity of the model.

Examples of Polynomials

A polynomial of degree 1 is of the form:

\[ f(X_1, X_2, \ldots, X_n) = w_0 + w_1 X_1 + w_2 X_2 + \ldots + w_n X_n \]

A polynomial of degree 2 is of the form:

\[ f(X_1, X_2, \ldots, X_n) = w_0 + w_1 X_1 + w_2 X_2 + \ldots + w_n X_n + w_{11} X_1 X_1 + w_{12} X_1 X_2 + w_{13} X_1 X_3 + \ldots + w_{ij} X_i X_j \]
Methodology

In the large majority of cases, a first degree polynomial is sufficient for generation of a relevant and robust model.

Using a higher degree of polynomial does not always guarantee better results than those obtained with a first degree polynomial. In addition, the higher the degree of polynomial you select:

- The more time needed to generate the corresponding model,
- The more time needed to apply the model to new datasets,
- The harder it is to interpret the results of modeling.

The selection of the degree of the polynomial depends on the nature of the data to be analyzed. The recommended method is to:

- First generate a model with a first order model. In the large majority of cases, this degree will be sufficient to guarantee a relevant and robust model.
- Test the results thus obtained with models of greater degree, if the performance of the first order model seems inadequate.

4.7.3 Validating the Model

Once the model has been generated, you must verify its validity by examining the performance indicators:

- The predictive power allows you to evaluate the explanatory power of the model, that is, its capacity to explain the target variable when applied to the training dataset. A perfect model possesses a predictive power of 1 and a completely random model possesses a predictive power of 0. No minimum threshold is required for the predictive power of a model. This depends upon the context of your work, that is, your domain of application, the nature of your data and your business issue. In some cases, a model with a predictive power as low as 0.1 may allow realization of a profit of several thousands dollars. In all cases, a positive predictive power indicates that the model generated will perform better than a random model.
- The prediction confidence defines the degree of robustness of the model, that is, its capacity to achieve the same explanatory power when applied to a new dataset. In other words, the degree of robustness corresponds to the predictive power of the model when applied to an application dataset. A model with a prediction confidence inferior to 0.95 must be considered with caution. The performance of such a model is very likely to vary between the training dataset and the application datasets.

To discover how these indicators are calculated, see the related topic on Predictive Power, Prediction Confidence and Model Graphs.

i Note

Validation of the model is a critically important phase in the overall process of data mining. Always be sure to assign significant importance to the values obtained for the predictive power and the prediction confidence of a model.
4.7.4 How to Obtain a Better Model

Obtaining a better model is achieved by:

- Improving the prediction confidence of the model, or
- Improving the predictive power of the model, or
- Improving both the predictive power and the prediction confidence of the model.

Several techniques allow you to improve these indicators:

- You can increase the degree of complexity of the model (polynomial degree).
- The following table presents other techniques.

<table>
<thead>
<tr>
<th>To improve...</th>
<th>You can...</th>
</tr>
</thead>
<tbody>
<tr>
<td>The predictive power of a model</td>
<td>○ Add variables to the training dataset</td>
</tr>
<tr>
<td></td>
<td>○ Use combinations of explanatory variables that seem relevant to you</td>
</tr>
<tr>
<td>The prediction confidence of a model</td>
<td>Add observations to the training dataset</td>
</tr>
</tbody>
</table>

For more information about improving the predictive power and the prediction confidence, see the related topic.

4.8 Performance Indicators

Two indicators specific to Automated Analytics allow you to evaluate the performance of a model.

- The predictive power, which is the quality indicator of the model,
- The prediction confidence, which is the robustness indicator.

Three other indicators, commonly used in data mining, are provided to assess a model:

- the GINI index,
- the K-S
- the AUC
Related Information

**Quality Indicator: Predictive Power** [page 39]
**Robustness Indicator: Prediction Confidence** [page 39]
**Predictive Power, Prediction Confidence and Model Graphs** [page 40]
**Advanced Users: Predictive Power for Continuous Targets** [page 41]
**Gini Index** [page 42]
**K-S** [page 42]
**AUC** [page 43]
**Error Indicators** [page 44]

### 4.8.1 Quality Indicator: Predictive Power

The predictive power of a model is the quality indicator of models generated using the application. This indicator corresponds to the proportion of information contained in the target variable that the explanatory variables are able to explain.

To improve the predictive power of a model, new variables may be added to the training dataset. Explanatory variables may also be combined.

**Example**

A model with a predictive power of:

- **"0.79"** is capable of explaining 79% of the information contained in the target variable using the explanatory variables contained in the dataset analyzed.
- **"1"** is a hypothetical perfect model, capable of explaining 100% of the target variable using the explanatory variables contained in the dataset analyzed. In practice, such a predictive power would generally indicate that an explanatory variable 100% correlated with the target variable was not excluded from the dataset analyzed.
- **"0"** is a purely random model.

### 4.8.2 Robustness Indicator: Prediction Confidence

The prediction confidence is the robustness indicator of the models generated using the application. It indicates the capacity of the model to achieve the same performance when it is applied to a new dataset exhibiting the same characteristics as the training dataset.

To improve the prediction confidence of a model, additional observation rows may be added to the training dataset.
Example

A model with a prediction confidence:

- Equal to or greater than "0.98" is very robust. It has a high capacity for generalization.
- Less than "0.95" must be considered with caution. Applying it to a new dataset will incur the risk of generating unreliable results.

4.8.3 Predictive Power, Prediction Confidence and Model Graphs

On the model graph plot:

- Of the estimation data set (default plot), the predictive power corresponds to "the area found between the curve of the model generated and that of the random model" divided by "the area found between the curve of the perfect model and that of the random model". As the curve of the generated model approaches the curve of the perfect model, the value of the predictive power approaches 1.
- Of the estimation, validation and test data sets (select the corresponding option from the list Data set, located below the plot), the prediction confidence corresponds to one minus "the area found between the curve of the estimation data set and that of the validation data set" divided by "the area found between the curve of the perfect model and that of the random model".

The following graph displays the predictive power and the prediction confidence:
4.8.4 Advanced Users: Predictive Power for Continuous Targets

The regression (continuous target case) uses the predictive power (KI) for model selection. If we suppose that we want to compute both the predictive power of a score variable \( rr_T \) with respect to its target \( T \) on a Validation dataset, then, we will refer to target categories by \( T_j, j = 1 \ldots B_T \). Thus, we note:

\[
U_j = \text{mean} \left( T_j \right), j = 1 \ldots B_T \\
f_j = \text{frequency} \left( T_j \right), j = 1 \ldots B_T 
\]

The target segments are given in decreasing natural order, so that \( u_1 > u_2 > \cdots > u_{B_T} \). Let \( \mu \) be the global mean of the target \( T \) on the Validation dataset.

We also refer to the score categories/segments by \( S_j \) for \( j = 1 \ldots B_S \) and note:

\[
m_j = \text{target mean} \left( S_j \right) \text{ for } j = 1 \ldots B_S \\
F_j = \text{frequency} \left( S_j \right), j = 1 \ldots B_S 
\]

The wizard curve is given by the cumulative profits of the target as a function of cumulative frequencies; it is defined by the following points:

\[
\left( \sum_{j = 1}^{b} \sum_{j = 1}^{b} f_{jj} (u_j - u) \right) \text{ for } b = 1 \ldots B_T 
\]

This curve is normalized such that its maximum is equal to 1.

The Validation curve is given by the cumulative score profits as a function of cumulative frequencies:

\[
\left( \sum_{j = 1}^{b} \sum_{j = 1}^{b} F_{jj} (m_j - u) \right) \text{ for } b = 1 \ldots B_S 
\]

As usual, the Predictive Power value is computed from the wizard and the Validation curve areas. For example, areas can be computed using a trapezoidal rule.

**Note**

The nominal target case can be viewed as a special case where the notion of the profit is the positive rate (which is equivalent to the binary target mean in this case).
4.8.5 Gini Index

The Gini statistic is a measure of predictive power based on the Lorenz curve. It is proportionate to the area between the random line and the model curve.

The Gini index is defined as the area under the Lorenz curve. The Gini index is the area between the ‘Trade-off’ curve and the obtained curve multiplied by 2. This is often pictured as the following chart:

![Gini Index Chart]

The horizontal axis ‘grows’ with the score and can be associated with 1-f.:

This is simply expressed using our notations as:

$$\text{GINI} = 2 \left( \frac{1}{2} - \int_{-\infty}^{+\infty} (1 - a(t)) d \left( (1 - a(t)) \ast P_G + \beta(t) \ast (1 - P_G) \right) \right)$$

$$\text{GINI} = 2 \left( 1 - P_G \right) \left( \frac{1}{2} - (1 - AUC) \right) = (1 - P_G)(2 AUC - 1)$$

Using these notations, we know that the Gini index of a random model is 0 and for a perfect model is $1 - P_G$.

Related Information

Lorenz Curves [page 48]

4.8.6 K-S

K-S is the Kolmogorov-Smirnov statistic applied here as a measure of deviation from uniform response rates across categories of a variable. Kolmogorov-Smirnov is a non-parametric, exact goodness-of-fit statistic based on the maximum deviation between the cumulative and empirical distribution functions.

In the case of a binary classification task, people are interested by the difference between the Lorenz curve for the good cases $1 - \alpha$, and the Lorenz curve for the bad cases $\beta$ when selecting an increasing ratio of population. These curves evolve from 0 to 1 together, and the K-S statistics is the maximum deviation between these two curves. For a perfect system, the K-S statistics is 1 and that for a random system, because of the equality between the two curves, the K-S statistics is 0.
Tip

The K-S is used to calculate the difference between two distributions in order to have an idea about the quality of a dataset.

Related Information

Lorenz Curves [page 48]

4.8.7 AUC

The Area Under the Roc Curve (AUC) statistic is a rank-based measure of model performance or predictive power calculated as the area under the Receiver Operating Characteristic (ROC) curve. For a simple scoring model with a binary target, this represents the observed probability of a signal (responder) observation having a higher score than a non-signal (non-responder) observation. For individual variables, ordering based on score is replaced by ordering based on the response probability for the variable’s categories (for example cluster ID or age range response rates).

The corresponding equation is:

\[
AUC = \int_{-\infty}^{+\infty} x(t) d(1 - \beta(t)) = \int_{0}^{1} x(y) dy
\]

So we have:

\[
AUC_{\text{perfect}} = 1
\]

\[
AUC_{\text{random}} = \int_{0}^{1} (y) dy = \frac{1}{2}
\]

One of the interests of the measure of AUC is its independence from the target distribution. For example, if we build another dataset and duplicate each good example twice, the AUC of the model will be the same.

Note

AUC has good properties for evaluating a binary classification system. It is widely used by statisticians.

Related Information

ROC [page 48]
4.8.8 Error Indicators

First, some basic notations:

- Target (response value): $y_i$
- Predictor (predictor response value): $\hat{y}_i$
- Residual: $r_i = y_i - \hat{y}_i$
- Error: $u_i = |y_i - \hat{y}_i| = |r_i|$
- Weight of the tested observation: $w_i$
- Total weight of the population: $W = \sum_{i=1}^{n} w_i$

- Target average: $\bar{y} = \frac{1}{W} \sum_{i=1}^{N} w_i y_i$
- Predictor average: $\bar{\hat{y}} = \frac{1}{W} \sum_{i=1}^{N} w_i \hat{y}_i$

Mean Absolute Error (L1)

Definition: mean of the absolute values of the differences between predictions and actual results, (City block distance or Manhattan distance)

Formula:

$$L1 = \frac{1}{W} \sum_{i=1}^{N} w_i |r_i|$$

Mean Square Error (L2)

Definition: square root of the mean of the quadratic errors (Euclidian distance or root mean squared error - RMSE)

Formula:

$$MSE = \frac{SSE_w}{W} = \frac{1}{W} \sum_{i=1}^{N} w_i u_i^2$$
Maximum Error (LInf)

Definition: maximum absolute difference between predicted and actual values (upper bound) (Chebyshev distance)

Formula:

\[ L_\infty = \max_i u_i \]

Error Mean

Definition: mean of the difference between predictions and actual values

Formula:

- Mean Percent Error (MPE):

\[ MPE = \frac{1}{W} \sum_{i=1}^{N} w_i \frac{y_i - \hat{y}_i}{y_i} \]

- Mean Absolute Percent Error (MAPE):

\[ MAPE = \frac{1}{W} \sum_{i=1}^{N} w_i \frac{|y_i - \hat{y}_i|}{y_i} \]

Error Standard Deviation

Definition: dispersion of errors around the actual result

Formula:

\[ \sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (u_i - \bar{u})^2} \]

where

\[ \bar{u} = \frac{1}{N} \sum_{i=1}^{N} u_i \]
Classification Rate

Definition: ratio between the number of correctly classified records and the total number of records

Formula:

$$\bar{u} = \frac{1}{N} \sum_{i=1}^{N} u_i$$

Determination Coefficient (R2)

Definition: ratio between the variability (sum of squares) of the prediction and the variability (sum of squares) of the data.

Formula:

$$SSR = \sum_{i=1}^{N} W_i (\mathcal{P}_u - \overline{\mathcal{P}})^2$$

$$SST = \sum_{i=1}^{N} w_i (y_i - \overline{y})^2$$

$$R^2 = \frac{SSR}{SST}$$

4.9 Profit Type

A profit type allows calculation of the profit that may be realized using the model. In general, a benefit is associated with the positive (or expected) values of the target variable and a cost is associated with the negative (or unexpected) values. For instance, in the context of a promotional mailing campaign, a person is associated with:

- A benefit for responding to the promotional mailing,
- A cost for not responding to the promotional mailing.

To visualize the profit that may be realized using a model generated by the application, you may use the following profit types:

- Detected profit,
- Lift profit,
Detected Profit

Detected profit is the profit type shown as the default. It allows examination of the percentage of observations belonging to the target category of the target variable, that is, the least frequent category, as a function of the proportion of observations selected from the entire dataset. Using this profit:

- The value "0" is assigned to observations that do not belong to the target category of the target variable,
- The value "1" (frequency of the target category of the target variable in the dataset) is assigned to observations that do belong to the target.

Lift Profit

Lift profit allows examination of the difference between a perfect model and a random model and between the model generated by the application and a random model. It represents the ratio between a model and the random model, that is the performance of a model compared to a model that would only allow to select observations at random from your database.

You can thus visualize how much better your model is compared with the random model.

Standardized Profit

Standardized profit allows examination of the contribution of the model generated by the application features relative to a model of random type, that is, in comparison with a model that would only allow selecting observations at random from your database.

This profit is used for the plots of variable details, which present the significance of each of the categories of a given variable with respect to the target variable.

Customized Profit

Customized profit allows you to define your own profit values, that is, to associate both a cost and a benefit to each value of the target variable. For instance, you can define:

- the cost of sending out a mailing as a negative value, for example "-5",
- the benefit brought in by the response to that mailing as a positive value, for example "20".
4.10 Advanced Model Curves

4.10.1 ROC

The ROC (Receiver Operating Characteristic) graph is derived from signal detection theory. It portrays how well a model discriminates in terms of the tradeoff between sensitivity and specificity, or, in effect, between correct and mistaken detection as the detection threshold is varied.

Sensitivity, which appears on the Y axis, is the proportion of CORRECTLY identified signals (true positives) found (out of all true positives on the validation dataset).

\[1 – \text{Specificity}\], which appears on the X axis, is the proportion of INCORRECT assignments to the signal class (false positives) incurred (out of all false positives on the validation dataset). (Specificity, as opposed to \[1 – \text{specificity}\], is the proportion of CORRECT assignments to the class of NON-SIGNALS – true negatives.)

4.10.2 Lorenz Curves

Lorenz "Good"

Lorenz 'Good' displays the cumulative proportion of missed signals (false negatives) accounted for by the records corresponding to the bottom x% of model scores.
The Y axis measures \([1 – sensitivity]\), that is \([1 – the\ proportion\ of\ true\ positives]\), which is equivalent to the proportion of missed signals or lost opportunity. Because the data are ordered from records predicted least likely to be signals on the left to records most likely to represent signals on the right, the slower the rise, the more sensitive the model in terms of detecting signals (or responders). The wizard line turns upward from the x-axis at the point corresponding to the proportion of non-signals in the validation dataset.

**Lorenz “Bad”**

Lorenz “Bad” displays the cumulative proportion of true negatives (specificity) accounted for by the bottom x% of model scores. Here, the faster the rise, the lower the frequency of erroneous detection.

![Performance](image)

### 4.10.3 Density Curves

The density curves display the density function of the variable score in the set of events (Curve Density “Good”) and in the set of non events (Curve Density “Bad”). These curves can also be viewed as the derivate of Lorenz curves (the density function is by definition the derivate of the cumulative density function).

The estimated density function in a bin or interval is equal to:

\[
\text{Number of Events in the Interval} \quad \frac{\text{Total number of Events}}{\text{Length of the interval}}
\]

The length of an interval is by definition its upper bound minus its lower bound.

**Density "Good"**

This curve displays the distribution of model scores for responders/signals.
Density "Bad"

This curve displays the distribution of model scores for non-responders/non-signals.

Density "All"

This curve displays both the curves Density "Good" and Density "Bad", thus allowing you to compare both distributions.
4.10.4 Risk Curves

**Good/Bad Odds**

The X-axis represents the risk score and the Y-axis represents the odds ratio value.

The odds ratio is equal to \((1-p)/p\), where \(p\) is the probability of risk.

**Probability of Risk**

The X-axis represents the risk score and the Y-axis represents the odds ratio value.

The probability of risk \(p\) is computed for each risk score bin this way: number of "Bad" divided by the number of records in the risk score bin.
Population Density

The density is computed according to the number of records in each risk score bin (20 by default).

Risk 'All'

All three curves are displayed in the same graph. Note that the y-axis of the probability curve is on the right hand side. The y-axis of the population density and the good/bad odds on the left.
5 Modeler - Regression/Classification

5.1 Regression/Classification Application Scenario Description

In this scenario, you are the Marketing Director of a large retail bank. You want to enhance efficiency and master your budget using modeling. The bank wants to offer a new financial product to its customers. Your project consists of launching a direct marketing campaign aimed at promoting this product. You have a large database of prospects at your disposal and a limited and closely monitored budget, and you are also subject to significant time constraints. In order to maximize the benefits of your campaign, your business issue consists of:

- Contacting those prospects most likely to be interested in the new financial product,
- Identifying the ideal number of prospects to contact out of the entire database.

Using the Modeler - Regression/Classification (formerly known as K2R) feature, you can rapidly develop an explanatory and predictive model at the least possible cost. This model allows you to respond to your business issue and accomplish your objectives.

Your Objective

Imagine the following case.

You are the Marketing Director of a large retail bank. This bank has decided to offer its customers a new high-end savings product. It prepares to launch an extensive direct marketing campaign to promote this new product to its prospects and customers.

The bank is experiencing heavy competition and senior management, sensitive to the stakes involved in launching this new financial product, wants the marketing campaign completed as soon as possible.

Your Means

A limited and closely monitored budget: The enterprise controls of the bank are rigorous, and the budget that has been allocated to you for this marketing campaign:

- Does not allow you to contact all of the bank’s prospects and customers,
- May not be exceeded.

The information at your disposal: The Marketing Department has a database for this campaign which contains the records of 1,000,000 prospects, identified by their principal characteristics, including:

- Age
You note that the database you have at hand is not ideal. In fact, the database contains:

- Incongruous data: The database contains alphanumeric information (such as "occupation" and "nationality") as well as numerical information (such as "age" and "unreconciled accounts").
- Redundant data: Some information in the database is redundant, such as "degree" and "education", or "degree" and "area of work".
  In the field of statistics, the term "correlated variables" is used to designate such data. In classical statistical analyses, correlated variables must be processed in a particular manner. An alternate solution is to designate only one of the two correlated variables for analysis.
  Since you have neither the statistical skills nor the means to handle this issue of correlation between variables, you decide to leave the database as it is.
- Missing data: Some information is missing from the database. To manage this lack of information, the Information Technology department used the following convention:
  - The symbol "?" means that an alphanumeric value (such as "occupation") is missing.
  - The value "99999" means that a numerical value (such as "age") is missing.
  Unfortunately, you have neither the time nor the necessary resources to perform a survey to fill in the missing information or to re-format the database.

The database available to you is stored in an RDBMS (relational database management system) residing on a UNIX server, maintained by the Information Technology department of the bank. The technical constraints of this information environment are determining factors in selecting potential data analysis tools.

**Your Approach**

By virtue of the critical stakes involved in this campaign, because of your limited budget and your inability to predict customers’ enthusiasm for the new product, you have chosen to minimize your risks by dividing the project into two steps:

- Test the marketing campaign on a sample of 50,000 individuals extracted from the prospects database of 1,000,000 people.
- Global launch of the marketing campaign using the entire contents of the prospects database.

The test phase of your marketing campaign allowed you to collect a sample of 50,000 individuals whose behavior with respect to this new product is known:

- 25% of the prospects showed themselves to be clearly interested. They chose to accept an invitation for a meeting with one of your sales channel agents,
- 75% of the prospects declined your invitation.

Your business issue consists of understanding the test results, by identifying the reasons which led certain individuals to respond favorably to your offer, and others to respond in the negative. Then you would be able to use the analytical model obtained to predict the behavior of each of the 1,000,000 prospects in your database. This would ensure that you optimized your marketing campaign by making the offer only to those individuals most likely to be interested.
The file containing the dataset used for the test was sent to you by the Information Technology department of the bank in the form of a flat file (.csv). This file corresponds to the sample file Census01.csv provided with the application. For more information, see the related topic about sample files.

Your Business Issue

Following the test phase of your campaign, your marketing database will contain:

- A list of 1,000,000 prospects.
- A list of 50,000 prospects, selected in a random manner during the test phase, and whose response to your test campaign is known. This sample, taken from your initial database, also contains missing values and correlated variables.

Your approach to the business issue consists of using the dataset in its present state, as a training dataset, in order to:

- Rapidly create an explanatory and predictive model.
- Next, apply this model to the entire database.

Using the model generated, you will be able to determine:

- How many individuals contained in your prospect database you should send your mailing to in order to maximize the profit/return on investment of your campaign.
- How to classify all of the individuals in your prospect database according to their interest (purchasing probability) in this new product. This interest is expressed as a score, or probability that a prospect will respond favorably to the campaign.
- What characterizes these individuals and what are their profiles? Validate the criteria (age, socio-occupational class, degree) that explain why a person expresses interest or not in the new financial product.
- How to simulate in real time, the likelihood of a single individual to respond favorably to a new offer, in particular, to allow the call center of your bank or a customer service agent to immediately know the level of interest that a prospect is likely to exhibit in this financial product.
- How to record this score in your prospect database, in order to be able to select sub-groups of the population for new campaigns at a later date.
- How to measure the quality and reliability (capacity of handling new individuals) of your model.

In order to allow you to better respond to these issues, you have access to several possible application solutions.

Your Solutions

To select the individuals to whom you will send a mailing, you have several possible solutions. You can use:

- A shotgun method.
- An intuitive method.
- A classical statistical method (for example: neural networks, Bayesian networks, logistic models, decision trees)
5.2 Scenario Solutions

5.2.1 Shotgun Solution Method

This method consists of performing no selection on your database, and sending out a mass mailing to every person recorded in your database. This solution guarantees that all persons likely to purchase your product are contacted.

On the other hand, the costs of this solution can be high, potentially far exceeding your budget, and it is seldom the solution applied. In addition, it runs the risk of saturating the prospects of your bank with inappropriate offers (spamming).

5.2.2 Intuitive Solution Method

This method consists of performing a selection that leans on your knowledge of your field, that is to say, you send your mailing to individuals selected in an intuitive manner from your database. This solution allows you to significantly reduce the cost of your marketing campaign and make it fit your budget.

This method is not optimal, because it does not allow you to:

- Control the real costs and return on investment of your marketing operation.
- Select which prospects to contact on a basis of real returns. It is true that you probably have a relatively good understanding of which individuals stand a good chance of becoming your customers some day. But optimizing your campaign means being able to identify those prospects that have every chance of becoming customers today as a result of the current marketing campaign.
- Discover new niche prospects that all your knowledge of the market had not previously allowed you to identify.
- Select a predefined number of individuals. Imagine that one of the constraints of your campaign consisted of contacting exactly 5,000 prospects. Your intuition may help you to select 2,400 of these. But how are you going to identify the remaining 2,600 prospects to be contacted? A purely random selection, thus completely non-optimized, might be your only solution.
5.2.3 Classical Statistical Solution Method

You may decide to use a classical statistical method to better manage the effectiveness of your campaign, and thus, your budget.

On the basis of the information that you have, a Data Mining expert could create predictive models. In other words, you could ask a statistical expert to create a mathematical model that would allow you to predict the probability of a given individual to respond to your marketing campaign, as a function of his profile.

To implement this method, the statistician must:

- Perform a detailed analysis of your test campaign,
- Prepare your database down to the smallest detail, specifically, encoding the different types of data in such a way that they can be used by the analytical tools he will apply,
- Test different types of algorithms (for example: neural networks, Bayesian networks, logistic models, decision trees) and select the one best suited to your business issue.

Typically, after a few weeks, the statistician will be able to associate a value with each individual in your database, indicating the probability of being interested or not interested in your marketing campaign.

This method presents significant constraints. You must:

- Ensure that your statistical expert, perhaps from a department external to the Marketing Department, is available for the scheduled period,
- Ensure that the cost for using this scarce resource will fit into your budget,
- Spend time explaining your domain-specific business issue to him,
- Spend time understanding the results that are provided.

5.2.4 Automated Analytics Solution Method

The simplicity and automatic nature of the application will allow you to perform the statistical analysis of your database yourself in a short amount of time.

Using the application, you will be able to create a model that allows you to:

- Determine which individuals have the highest probability (score) of being interested in your marketing campaign (predictive modeling). You may then apply the model to your entire database.
- Break out the determining factors that describe the phenomenon that you hope to model, that is, the fact of being "interested" or "not interested" in the new financial product of the bank (descriptive modeling).

The profit curve, an important validation and control tool, allows you to compare the performance of models generated using the application features with that of a hypothetical random model and that of a hypothetical perfect model. At the same time, it also allows you to determine the optimal number of persons to contact to maximize the profit generated by your campaign. The application also provides you with indicators of the quality of the model you generate (predictive power), and its capacity to generalize and remain relevant to new datasets (prediction confidence).

The application provides you with the means to customize your direct marketing campaign with respect to your different customer profiles, increasing your powers of persuasion.
5.3 Introduction to Sample Files

This guide is accompanied by the following sample data files:

- A data file `Census01.csv`.
- The corresponding description file `desc_census.csv`.

These files allow you to evaluate Automated Analytics features and take your first steps in using it.

`Census01.csv` is the sample data file that you will use to follow the scenarios for Regression/Classification and Segmentation/Clustering. This file is an excerpt from the American Census Bureau database, completed in 1994 by Barry Becker.

**Note**

For more information about the American Census Bureau, see [http://www.census.gov](http://www.census.gov).

This file presents the data on 48,842 individual Americans, of at least 17 years of age. Each individual is characterized by 15 data items. These data, or variables, are described in the following table.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Example of Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>Age of individuals</td>
<td>Any numerical value greater than 17</td>
</tr>
<tr>
<td>workclass</td>
<td>Employer category of individuals</td>
<td>Private, Self-employed-not-inc</td>
</tr>
<tr>
<td>fnlwgt</td>
<td>Weight variable, allowing each individual to represent a certain percentage of the population</td>
<td>Any numerical value, such as &quot;0&quot;, &quot;2341&quot; or 205019.</td>
</tr>
<tr>
<td>education</td>
<td>Level of study, represented by a schooling level, or by the title of the degree earned</td>
<td>11th, Bachelors</td>
</tr>
<tr>
<td>education-num</td>
<td>Number of years of study, represented by a numerical value</td>
<td>A numerical value between 1 and 16</td>
</tr>
<tr>
<td>marital-status</td>
<td>Marital status</td>
<td>Divorced, Never-married</td>
</tr>
<tr>
<td>occupation</td>
<td>Job classification</td>
<td>Sales, Handlers-cleaners</td>
</tr>
<tr>
<td>relationship</td>
<td>Position in family</td>
<td>Husband, Wife</td>
</tr>
<tr>
<td>race</td>
<td>Ethnicity</td>
<td>White, Black</td>
</tr>
<tr>
<td>sex</td>
<td>Gender</td>
<td>Male, Female</td>
</tr>
<tr>
<td>capital-gain</td>
<td>Annual capital gains</td>
<td>Any numerical value</td>
</tr>
</tbody>
</table>
### Variable Description Example of Values

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Example of Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>capital-loss</td>
<td>Annual capital losses</td>
<td>Any numerical value</td>
</tr>
<tr>
<td>native country</td>
<td>Country of origin</td>
<td>• United States</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• France</td>
</tr>
</tbody>
</table>
| class          | Variable indicating whether or not the salary of the individual is greater or less than $50,000 | • “1” if the individual has a salary of greater than $50,000
|                |                           | • “0” if the individual has a salary of less than $50,000 |

**i**  Note

In order to avoid complicating the application scenarios, the variable fnlwgt is used as a regular explanatory variable in these scenarios, and not as a weight variable.

### 5.4  SAP BusinessObjects Predictive Analytics

To accomplish the scenario, you will use the Automated Analytics toolset of SAP BusinessObjects Predictive Analytics.

#### 5.4.1 To Start SAP BusinessObjects Predictive Analytics

1. Select [Start] ➤ [Programs] ➤ [SAP Business Intelligence] ➤ [SAP BusinessObjects Predictive Analytics Desktop] ➤ [SAP BusinessObjects Predictive Analytics]

   The SAP BusinessObjects Predictive Analytics start panel appears.
2. Click the feature you want to use.

#### 5.4.2 Editing the Options

- To edit the application options:

1. In the [File] menu, click [Preferences]...

   The window [Edit Options...] appears.
2. The following options can be modified for Classification/Regression:

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Category</td>
<td>Options</td>
</tr>
</tbody>
</table>
### Option Description

| General     | ○ Country <br>○ Language <br>○ Message Level <br>○ Log Maximum Size <br>○ Message Level for Strange Values <br>○ Display the Parameter Tree <br>○ Number of Store in the History <br>○ Always Exit without Prompt <br>○ Include Test in Default Cutting Strategy <br>○ SQL Statement Separator |
| Stores      | ○ Default Store for Apply-in Data Set <br>○ Default Store for Apply-out Data Set <br>○ Default Store to Save Models |
| Metadata Repository | ○ Store the Metadata in the Same Place as the Data <br>○ Store the Metadata in a Single Place <br>○ Edit Variable Pool Content |
| Graphic     | ○ Graphic theme <br>○ Font Size <br>○ Profit Curve Points <br>○ Bar Count Displayed <br>○ Display 3D Chart <br>○ Disable Double Buffering <br>○ Optimize for Remote Display |
| Report      | ○ Report Custom Banner <br>○ Style Sheet Repository <br>○ Active Style Sheet |
| Geolocation | ○ Geographic Information System Protocol |

## 5.4.2.1 Customizing Style Sheets

SAP BusinessObjects Predictive Analytics offers the possibility to customize the generated reports. The default style sheet, called *SAP BusinessObjects Predictive Analytics (default)*, cannot be modified. You have to create your own style sheets to modify the settings.

**Note**

Before you create, load or save a style sheet, you must first select a *Style Sheet Repository* folder.

### Related Information

- To Create a New Style Sheet [page 62]
- To Delete a Style Sheet [page 62]
- To Edit the General Settings [page 62]
5.4.2.1.1  To Create a New Style Sheet

1. In the field Folder, click the button 📲 (Browse).
2. Select a folder.
   This folder is your style sheets repository.
3. Click the button ✦ (Add).
   A new style sheet has been created.
4. Click the button 📖.
   The panel Report Style Sheet Editor opens.
5. In the field Style Sheet Name, enter a name for the new style sheet.
   The extension .krs is automatically added.

ℹ️ Note
You can duplicate a style sheet by changing the name of your style sheet. The previous one is not deleted.

5.4.2.1.2  To Delete a Style Sheet

1. Select one of the displayed style sheets.
2. Click the button └ (Remove).

ℹ️ Note
The style sheet is not only deleted from the list but also from the data source.

5.4.2.1.3  To Edit the General Settings

Select settings:
5.4.2.1.4 To Edit the Charts Settings

Select settings:

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Settings...</td>
<td>Options...</td>
</tr>
<tr>
<td>Chart Colors</td>
<td>modify the chart colors</td>
</tr>
</tbody>
</table>
| Default Chart Bars Orientation | ○ horizontal  
                             | ○ vertical                           |

**Note**

It is possible to set another default orientation for specific report items.

5.4.2.1.5 To Edit Report Items

1. Set the properties of your choice.
<table>
<thead>
<tr>
<th>Properties...</th>
<th>Functions...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Displayed as</td>
<td>name of the label</td>
</tr>
<tr>
<td>View Type</td>
<td>choose between Tabular, HTML and Graphical. The last one is only available if the report item can be displayed as a graph.</td>
</tr>
<tr>
<td>Chart Type</td>
<td>select one of the proposed chart types. Note that this option is only available for report items of the view type Graphical.</td>
</tr>
<tr>
<td>Switch Bar Orientation</td>
<td>this option allows having another bar orientation as the default one for a specific report item</td>
</tr>
<tr>
<td>Sort by/Sort Order</td>
<td>you can select a column to sort by and choose between an ascending or a descending order</td>
</tr>
<tr>
<td>Visibility</td>
<td>you can hide columns of a report item or even menu items. Note that at least one column of a report item must remain visible.</td>
</tr>
</tbody>
</table>

2. Click Save to validate.
   A window opens, indicating that your style sheet has been successfully saved.
3. Click OK.

### 5.4.2.1.6 To Apply the New Style Sheet to the Generated Reports

1. In the panel Report, select the new style sheet.
2. Click OK.
   A window opens, indicating that you have to restart the modeling assistant to take the edited options into account.
3. Click OK.
   When training a model, all the generated reports (the learn/excel/statistical reports) are now customized.

### 5.4.2.2 Defining a Metadata Repository

The metadata repository allows you to specify the location where the metadata should be stored.

1. Choose between storing the metadata in the same place as the data or in a single place by checking the option of your choice.
2. In the list Data Type, select the type of data you want to access. For some types of data, you will need a specific license.
3. Use the Browse button corresponding to the Folder field to select the folder or database containing the data. In case of a protected database, you will need to enter the user name and the password in the fields User and Password.
4. Click the button *Edit Variable Pool Content* to edit the parameters of the variables stored in the variable pool.

5. Click *OK* to validate.

### 5.5 Creating a Model Using Modeler

Data modeling with the application is subdivided into four broadly defined stages:
- Defining the Modeling Parameters
- Generation and Validation of the Model
- Analysis and Understanding of the Analytical Results
- Using a Generated Model

### 5.5.1 Step 1 - Defining the Modeling Parameters

To respond to your business issue you want to:
- Identify and understand the factors that determine whether a prospect reacts positively or negatively to your marketing campaign.
- Thereby be able to predict the behavior of new prospects with respect to your campaign.

The Regression/Classification feature (formerly known as K2R) allows you to create explanatory and predictive models.

The first step in the modeling process consists of defining the modeling parameters:
- Select a data source to be used as training dataset.
- Describe the dataset selected.
- Select the variables: the target variables, the explanatory variables and possibly a weight variable.
- Check the modeling parameters.
- Setting the Advanced Parameters (degree, target key, variable auto-selection and correlations). This step is optional.

### Related Information

- [Setting the Advanced Parameters](page 84)
- [Selecting a Data Source](page 66)
- [Describing the Data Selected](page 66)
5.5.1.1 Selecting a Data Source

Use the file Census01.csv as a training dataset.

This file represents the sample that you had extracted from your database and used for the test phase of your direct marketing campaign. As specified in your test plan, this file contains data concerning 50,000 prospects, for whom you now know the behavior with respect to the new financial product:

- 25% of the prospects showed themselves to be clearly interested. They chose to accept an invitation for a meeting with one of your sales channel agents,
- 75% of the prospects declined your invitation.

In this file, you created a new variable Class, which corresponds to the reaction of prospects contacted during the test. You assigned:

- The value "1" to those prospects who responded positively to your invitation,
- The value "0" to those prospects who responded negatively to your invitation.

To select a data source:

1. On the screen Select a Data Source, select the data source format to be used (Text files, Data Base, ...).
2. Click the Browse button.

   In the Data Selection dialog, browse to the <Installation Path>/Samples folder.

   **Note**

   Depending on your environment, the Samples folder may or may not appear directly at the root of the list of folders. If you selected the default settings during the installation process, you will find the Samples folder located in C:Program Files/SAP BusinessObjects Predictive Analytics/Desktop <version number>/Automated/

3. Double-click the Samples folder, then the Census folder.
4. Select the file Census01.csv, then click OK.
5. The name of the file appears in the Data Set field.
6. Click Next.

   The screen Data Description appears. Now you are ready to describe the data selected.

5.5.1.2 Describing the Data Selected

For this scenario:

- Select Text Files as the Data Type.
- Use the file Desc_Census01.csv as the description file for the Census01.csv data file.

To Select a Description File:

1. On the screen Data Description, click the button Open Description.

   The Load a Description window opens.
2. Select the type of your description file.
3. In the Folder field, select the folder where the description file is located with the Browse button. The folder selected by default is the same as the one you selected on the screen Select a Data Source.
4. In the File Name field, select the file containing the dataset description with the Browse button.

⚠️ Caution

When the space used for model training contains a physical variable named $KxIndex$, it is not possible to use a description file without any key for the described space.

When the space used for model training does not contain a physical variable named $KxIndex$, it is not possible to use a description file including a description about a $KxIndex$ variable since it does not exist in current space.

5. Click OK.
   The window Load a Description closes and the description is displayed on the screen Data Description.
6. Click Next.

### 5.5.1.2.1 Why Describe the Data Selected?

In order for the application to interpret and analyze your data, the data must be described. To put it another way, the description file must specify the nature of each variable, determining the storage format and type:

- Storage format: number (number), character string (string), date and time (datetime) or date (date).

   ➖➖ Note

   When a variable is declared as date or datetime, the date coder feature automatically extracts date information from this variable such as the day of the month, the year, the quarter and so on. Additional variables containing this information are created during the model generation and are used as input variables for the model.
   The date coder feature is disabled for Time Series.

- Type: continuous, nominal, ordinal or textual.

For more information about data description, see the related topics.

### Related Information

- Variables [page 28]
- Variable Storage Formats [page 30]
5.5.1.2.2 How to Describe Selected Variables

To describe your data, you can:

- Either use an existing description file, that is, taken from your information system or saved from a previous use of the application,
- Or create a description file using the Analyze option. In this case, it is important that you validate the description file obtained. You can save this file for later re-use. If you name the description file KxDoc_<SourceFileName>, it will be automatically loaded when clicking the Analyze button.

Caution

The description file obtained using the Analyze option results from the analysis of the first 100 lines of the initial data file. In order to avoid all bias, we encourage you to mix up your data set before performing this analysis.

Each variable is described by the fields detailed in the following table:

<table>
<thead>
<tr>
<th>The Field</th>
<th>Gives information on…</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>the variable name (which cannot be modified)</td>
</tr>
<tr>
<td>Storage</td>
<td>the type of values stored in this variable:</td>
</tr>
<tr>
<td></td>
<td>• <strong>Number</strong>: the variable contains only “computable” numbers (be careful a telephone number, or an account number should not be considered numbers)</td>
</tr>
<tr>
<td></td>
<td>• <strong>String</strong>: the variable contains character strings</td>
</tr>
<tr>
<td></td>
<td>• <strong>Datetime</strong>: the variable contains date and time stamps</td>
</tr>
<tr>
<td></td>
<td>• <strong>Date</strong>: the variable contains dates</td>
</tr>
<tr>
<td>Value</td>
<td>the value type of the variable:</td>
</tr>
<tr>
<td></td>
<td>• <strong>Continuous</strong>: a numeric variable from which mean, variance, etc. can be computed</td>
</tr>
<tr>
<td></td>
<td>• <strong>Nominal</strong>: categorical variable which is the only possible value for a string</td>
</tr>
<tr>
<td></td>
<td>• <strong>Ordinal</strong>: discrete numeric variable where the relative order is important</td>
</tr>
<tr>
<td></td>
<td>• <strong>Textual</strong>: textual variable containing phrases, sentences or complete texts</td>
</tr>
</tbody>
</table>

Caution

When creating a text coding model, if there is not at least one textual variable, you will not be able to go to the next panel.

Key

whether this variable is the key variable or identifier for the record:

- 0 the variable is not an identifier;
- 1 primary identifier;
- 2 secondary identifier…
The Field... | Gives information on...
---|---
**Order** | whether this variable represents a natural order. (0: the variable does not repre­
sent a natural order; 1: the variable represents a natural order). If the value is set at
1, the variable is used in SQL expressions in an “order by” condition.
There must be at least one variable set as Order in the Event data source.

**Caution**
If the data source is a file and the variable stated as a natural order is not ac­
tually ordered, an error message will be displayed before model checking or
model generation.

| **Missing** | the string used in the data description file to represent missing values (e.g. “999”
or “#Empty” - without the quotes) |
| **Group** | Name of the group to which the variable belongs. You can assign a variable to one
group only. All variables of a group convey the same information. Thus different
groups cannot be crossed when the model has an order of complexity over 1. |
| **Description** | an additional description label for the variable |
| **Structure** | this option allows you to define your own variable structure, which means to define
the variables categories grouping. |

5.5.1.2.3 **Viewing the Data**

To help you validate the description when using the Analyze option, you can display the first hundred lines of
your data set.

1. Click the button **View Data**.
   A new window opens displaying the data set top lines:
2. In the field **First Row Index**, enter the number of the first row you want to display.
3. In the field **Last Row Index**, enter the number of the last row you want to display.
4. Click the **Refresh** button to see the selected rows.

5.5.1.2.4 **A Comment about Database Keys**

For data and performance management purposes, the data set to be analyzed must contain a variable that
serves as a key variable. Two cases should be considered:

- If the initial data set does not contain a key variable, a variable index \texttt{KxIndex} is automatically generated
  by the application. This will correspond to the row number of the processed data.
- If the file contains one or more key variables, they are not recognized automatically. You must specify them
  manually in the data description.
To Specify that a Variable is a Key

1. In the **Key** column, click the box corresponding to the row of the key variable.
2. Type in the value "1" to define this as a key variable.

5.5.1.2.5 Defining a Variable Structure

There are three ways to define a variable structure:

- by first extracting the categories from the variable statistics, then editing or validating the suggested structure.
- by importing the structure from an existing model.
- by building a new structure from scratch.

The option Optimal Grouping allows you to let Data Encoding group together the categories groups defined in the variable structure if they bring the same information.

The last column of the description table indicates the state of the structure of each variable. The following table lists the possible states of a variable structure.

<table>
<thead>
<tr>
<th>Icon</th>
<th>State</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Icon" /></td>
<td>undefined</td>
<td>Data Encoding will automatically determine the categories grouping depending on their interaction with the target variable.</td>
</tr>
<tr>
<td><img src="image2" alt="Icon" /></td>
<td>non-editable</td>
<td>The structure for an ordinal string variable cannot be modified.</td>
</tr>
<tr>
<td><img src="image3" alt="Icon" /></td>
<td>defined by extraction from the variable statistics</td>
<td>The user must open and validate the variable structure.</td>
</tr>
<tr>
<td><img src="image4" alt="Icon" /></td>
<td>defined by the user or imported from an existing model</td>
<td></td>
</tr>
</tbody>
</table>

**Note**

A translation of the variable categories has no influence on the variable structure, which has to be set according to the original values of the variable.

5.5.1.2.5.1 To Extract a Variable Structure

1. In the **Data Description** pane, select the variables for which you want to extract the structure.
2. Go to the **Structures** tab of the ribbon, the available options are separated in two parts: **Edit** and **Extract**.
3. Click the option **From Statistics** in the **Extract** part.

A progression bar is displayed while the structure is being extracted. Once the extraction is done, the icons corresponding to the selected variables change, indicating that the operation was a success and allowing you to easily identify them. You can then modify the variable structure as you need.
5.5.1.2.5.2 To Import the Variable Structure from a Model

1. Select the variables for which you want to extract the structure.
2. Go to the Structures tab of the ribbon, the available options are separated in two parts: Edit and Extract.
3. Select the option From Model and choose the desired option.
   The window Loading Model opens.
4. In the list Data Type, select the type of store the model is saved in.
5. Use the Browse button located next to the Folder field to select the folder or database containing the model.
6. In the displayed models list, select the model from which you want to extract the variable structure.
7. Click OK.
8. In the list Target from Loaded Model, select the target of the model.
   The variables you have selected are displayed in a list with the corresponding variables from the loaded model. You can add or remove variables from this list and view the model variables structure (see the related topics).
9. Once all the variables for which you want to import the structure from the model are displayed in the list, Click OK.
   The selection window closes and the structure state changes.

Related Information

To Add a Variable to the List of Variables [page 71]
To Remove a Variable from the List of Variables [page 71]
To View a Variable Structure Defined in the Loaded Model [page 72]

5.5.1.2.5.3 To Add a Variable to the List of Variables

1. In the list Variable from Loaded Model, select the variable you want to add to the list of variables for which the structure will be imported.
2. Click the Add button.
   ○ The variable appears in the list.

5.5.1.2.5.4 To Remove a Variable from the List of Variables

1. In the list located in the lower part of the panel, select the variable for which you do not want to import the structure.
2. Click the Remove button.

The variable is removed from this list and added to the list Variable from Loaded Model.

5.5.1.2.5.5 To View a Variable Structure Defined in the Loaded Model

1. If the variable has not been added yet to the list of variables located on the lower part of the panel:
   a. In the list Variable from Loaded Model, select the variable for which you want to see the structure defined in the model.
   b. Click the View button.

The variable structure opens in a new window.

2. If the variable has already been added to the list of variables located on the lower part of the panel, double-click the variable for which you want to see the structure defined in the model.

5.5.1.2.5.6 To Create or Modify a Variable Structure

1. Double-click the Structure icon corresponding to the variable for which you want to edit or create the structure.

The edit window opens. If the structure had been extracted from the variable statistics or from a model, the fields are already filled.

2. For details on how to use the structure editor, see the topics for the type of variable structure.

Related Information

Structure for a Continuous Variable [page 72]
Structure for an Ordinal Variable [page 74]
Structure for a Textual Variable [page 74]
Structure for a Nominal Variable [page 74]

5.5.1.2.5.6.1 Structure for a Continuous Variable

The structure for a continuous variable is defined by several intervals each made of:

- a lower bound ([ ] ) that can be either open or closed,
- a minimum value (Minimum),
- a maximum value (Maximum).
• a higher bound ([ ] ) that can be either open or closed.

All intervals must be adjoining: there can be no gap or overlap between two intervals.

The option Add Missing allows you to indicate with which interval the missing values should be grouped.

The option Include Smaller Data allows you to include to the first interval any value smaller than its lower bound. In the same way, the option Include Higher Data allows you to include to the last interval any value higher than its higher bound.

**To Create an Interval**

1. Click the Add button. The edit window opens.
2. Select the lower bound type by clicking the ] button.
3. Enter the minimum value for the interval in the left text field.
4. Enter the maximum value for the interval in the right text field.
5. Select the higher bound type by clicking the [ button.
6. Check the option Add Missing if the missing values must be grouped with this interval.
7. Click the Yes button to validate your interval.

**To Split an Interval**

1. Select the interval to split.
2. Click the Split button. The selected interval is automatically split into two equal intervals.

**To Merge an Interval**

1. Select the intervals to merge. You can only select adjoining intervals.
2. Click the Merge button.

**To Delete an Interval**

1. Select one or more intervals. You can only select adjoining intervals.
2. Click the Remove button. The previous and next intervals are extended to include the values previously contained in the deleted intervals, so that no gap is left between intervals.
5.5.1.2.5.6.2 Structure for an Ordinal Variable

The structure for an ordinal variable is similar to that of a continuous variable with the exception of the bounds which are always closed and cannot be modified.

⚠️ Caution

The structure for an ordinal string variable cannot be edited.

5.5.1.2.5.6.3 Structure for a Textual Variable

The structure for a textual variable cannot be edited.

5.5.1.2.5.6.4 Structure for a Nominal Variable

The structure for a nominal variable is made of groups containing the variable categories.

To Create a New Category Group

1. In the list **Category Edition**, select the categories you want to add in a new group. Use the Ctrl key to select several categories.
2. Click the button **Add New Group**. A group containing the selected categories is created in the list **Group Structure**.

To Include Missing Values in a Group

1. In the list **Group Structure**, select the group in which you want to add the missing values.
2. Click the button **Add Missing** located under the list **Category Edition**. The KxMissing category, which represents the missing values, is added to the selected group and the button **Add Missing** is deactivated. As any category, the KxMissing category can only belong to one group at a time.

To Create a New Category

- In the field right of the button **New Category**, enter the name of the category to add.
- Click the button **New Category**. The category is created in the list **Category Edition**.
To Add Categories to a Group

- In the list Category Edition, select the category (or categories) to add to a group.
- In the list Group Structure, select the group in which you want to add the selected categories.
- Click the button Add Category.

To Delete a Group

- In the list Group Structure, select the group to delete.
- Click the button Remove Group. All the categories belonging to this group are re-added to the list Category Edition.

To Remove a Category from a Group

- In the list Group Structure, select the category or categories you want to remove from the group.
- Click the button Remove Category. The selected categories are removed from the group and re-added to the list Category Edition.

5.5.1.2.6 Working Without any Defined Structure

If you let the structure as undefined, the application uses consistent coder to automatically determine the categories grouping depending on their interaction with the target variable. You can configure two parameters in this case:

- The band count for continuous variables
- Modeler - Data Encoding optimal grouping for all variables.

5.5.1.2.6.1 Band Count for Continuous Variables

When you work with no defined structure, you can set the band count for continuous variables. The allowed values for this parameter are between 1 and 20.

The population is thus divided into as many segments of similar size. These segments are used to build descriptive statistics, particularly the distribution of target variables for each segment, which affects the coding of the variable with respect to target variables.

The band count has an influence on the calculation of the predictive power: the more there are segments, the more accurate is the calculation of the predictive power for the explanatory variable. However, this influence is very small.
To Edit the Band Count for Continuous Variables

- Click the row corresponding to the continuous variable to be edited.
- Go to the Structures tab of the ribbon.
- Click Edit User Band Count. The Set Band Count screen displays.

<table>
<thead>
<tr>
<th>If you want to...</th>
<th>Then...</th>
</tr>
</thead>
</table>
| modify the band count for all the continuous variables of the model | ○ Type in the desired band count in the field at the bottom of the panel.  
○ Click Set the Same Band Count for All Variables.  
○ Click OK. |
| modify the band count for the variable being edited | ○ Type in the desired band count in the column Band Count at the top of the panel.  
○ Click OK. |

5.5.1.2.6.2 Optimal Grouping for All Variables

When working with a defined structure, if you want to keep your categories as they are defined for the model building, you must disable this option.

If not or if you work with no defined structure, Optimal Grouping allows in a large number of cases to increase the prediction confidence of the model with a minimal loss of predictive power. Where possible, similar adjacent segments are gathered to reduce artifacts between the estimation and validation datasets.

To Enable Modeler - Data Encoding Optimal Grouping for All Variables

- Right click the row corresponding to the variable to be edited.
- Select Define Structure.
- Select Optimal Grouping, in such a way that the option is checked.

5.5.1.3 Filtering the Data Set

In order to accelerate the learn process and to optimize the resulting model, you can apply a filter to your data set.

**Note**

For this scenario, do not use the filtering option.
5.5.1.3.1 To Filter a Data Set

1. Check the option *Add a Filter in Data Set*.
2. Click *Next*.

5.5.1.3.2 To Add a Condition

1. Click the button *Add a Condition*. The window *Define a Condition* opens.

2. Choose a variable in the first list.
3. Choose an operator in the second list.
4. Indicate a value in the third list:
   - For a variable with number storage, type a value.
   - For a variable with string storage, choose a variable in the list. If the list is empty click the button to extract the variable categories.
5. Click *OK*.

*i* Note

You can edit a condition by double-clicking it.

5.5.1.3.3 To Add a Logical Conjunction

Click the button *Add Logic 'And'* or the button *Add Logic 'Or'*.

You can change a conjunction by double-clicking it.
5.5.1.3.4 To Change the Order

You can change the order of the nodes to accelerate the filtering process by setting the conditions with the highest probability to be false at the top of the list.
1. Select the node you want to move up or down.
2. Use the buttons ↑ and ↓ to change its position in the filter.

5.5.1.3.5 To Delete a Node

1. Select the node you want to delete.
2. Click the button *Remove Selected Node*.

5.5.1.3.6 To Display the Filtered Data Set

You can visualize the data set that you will obtain after the application of the filter.
Click the button *View Data*.
A pop-up window opens.

5.5.1.3.7 To Save a Filter

You can save the filter you created to be able to reuse it at a later moment without being obliged to recreate the same conditions.
1. Click the button *Save Filter*.
   A pop-up window is displayed.
2. In the list *Data Type*, select the format in which you want to save the filter.
3. Use the *Browse* button located on the right of the *Folder* field to select the folder or database where you want to save the filter.
4. In the *File Name* field, enter the name of the file or table in which you want to save the filter.
5. Click the *OK* button.

5.5.1.3.8 To Load an Existing Filter

To apply a filter to the data set, you can use a file created during a previous use of the data set in Automated Analytics.
1. Click the button **Load Existing Filter**. A pop-up window is displayed.

2. Use the list **Data Type** to select the format of the filter.

3. Use the **Browse** button located on the right of the **Folder** field to select the folder or the database in which the filter is stored.

4. Use the **Browse** button located on the right of the **File** field to select the file or the table containing the filter.

5. Click the **OK** button.

### 5.5.1.4 Translating the Variable Categories

You can translate the categories of a nominal variable, save the translation or load an existing translation. This translation has no influence on the variable structure, which has to be set according to the original values of the variable.

**Note**

The variable "Target Key", which is used in the advanced settings, does not take into account the translation when displaying the possible values of this variable.

### Related Information

- To Translate the Variable Categories [page 79]
- To Save the Categories Translation [page 80]
- To Load an Existing Translation File [page 80]

### 5.5.1.4.1 To Translate the Variable Categories

1. Click a nominal variable to translate its categories.

2. Go to the **Edition** tab of the ribbon and click the option **Translate Categories**. A new window appears.

3. Choose into which languages you want to translate. By default, the language of the user interface is displayed as a column.

4. Click the **button to extract the variable categories from the data set.

5. Translate the categories.

**Note**

You do not need to fill all fields.
6. Click the OK button.

### 5.5.1.4.2 To Save the Categories Translation

1. After translating the variable categories, click the Save button.
2. Choose a Data Type.
3. Select a Folder.
4. Enter a Name for the file or table.
5. Click the OK button.

### 5.5.1.4.3 To Load an Existing Translation File

1. Click a nominal variable to translate its categories.
2. Go to the Edition tab of the ribbon and click the option Translate Categories.
   A new window appears.
3. Click the Load button.
4. Select the format of the translation in the list Data Type.
5. Use the Browse button located on the right of the Folder field to select the folder or the database in which
   the description is stored.
6. Use the Browse button located on the right of the field Table or File to select the file or the table containing
   the description.
7. Click the OK button.
8. Click the button Update to refresh the display of the categories.
9. If the list of columns is not named correctly, use the Advanced Settings to set a header line and
   update again.
10. Map the language names with those from the loaded translation, by clicking the categories and choosing
    the corresponding language in the contextual menu.
11. Click the OK button.

### 5.5.1.5 Selecting Variables

Once the training dataset and its description have been entered, you must select the following variables:

- One or more Target Variables
- Possibly a Weight Variable
- The Explanatory Variables
Related Information

Variables [page 28]

5.5.1.5.1  Target Variables

For this scenario, select the variable Class as your target variable, that is, the variable that indicates the probability of an individual responding in a positive or negative manner to your campaign.

1. On the screen Selecting Variables, in the section Explanatory variables selected (left hand side), select the variables you want to use as Target Variables.

   Note
   On the screen Selecting Variables, variables are presented in the same order as that in which they appear in the table of data. To sort them alphabetically, select the option Alphabetic sort, presented beneath each of the variables list.

2. Click the button > located on the left of the screen section Target Variables (upper right hand side).

   The variable moves to the screen section Target Variables.

   You can also select a variable in the screen section Target Variables and click the button < to move the variables back to the screen section Explanatory variables selected.

5.5.1.5.2  Weight Variable

Selecting a Weight Variable enables to set the Weight Quantum option available in the Advanced Model Parameters.

For this scenario, do not select a weight variable.

1. On the screen Selecting Variables, in the section Explanatory variables selected (left hand side), select the variables you want to use as Weight Variables.

   Note
   On the screen Selecting Variables, variables are presented in the same order as that in which they appear in the table of data. To sort them alphabetically, select the option Alphabetic sort, presented beneath each of the variables list.

2. Click the button > located on the left of the screen section Weight Variables (middle right hand side).

   The variable moves to the screen section Weight Variables.

   You can also select a variable in the screen section Weight Variables and click the button < to move the variables back to the screen section Explanatory variables selected.
5.5.1.5.3  Explanatory Variables

By default, and with the exception of key variables (such as KxIndex), all variables contained in your data set are taken into consideration for generation of the model. You may exclude some of these variables.

For the first analysis of your data set, we recommend that you retain all variables. It is particularly important to retain even the variables that seem to have no impact on the target variable. If indeed these variables have no impact on the target variable, the model will confirm this. In the contrary case, the model will allow you to recognize previously unidentified correlations between these variables and the target variable.

Depending on the results obtained from the first analysis, which included all of the variables of the data set, you can generate a second model by excluding the variables too closely correlated with the target variable.

For this Scenario:

- Exclude the variable KxIndex, as this is a key variable. Since the initial data set does not contain a key variable, the application generated KxIndex automatically.
- Retain all the other variables.

1. To exclude some variables from data analysis, on the screen Selecting Variables, in the section Explanatory Variables Selected (left hand side), select the variable to be excluded.

   **Note**
   On the screen Selecting Variables, variables are presented in the same order as that in which they appear in the table of data. To sort them alphabetically, select the option Alphabetic sort, presented beneath each of the variables list.

2. Click the button > located on the left of the screen section Excluded Variables (lower right hand side). The variable moves to the screen section Excluded Variables. Also, select a variable in the screen section Excluded Variables and click the button < to move the variables back to the screen section Explanatory Variables Selected.

   By default, any variable defined as a key is put in the Excluded Variables. However, the user has the possibility to move a key variable in the Explanatory Variables Selected if he wants this variable to have this role.

3. Click Next.

   The screen Parameters of the Model appears.

5.5.1.6  Checking Modeling Parameters

The screen Summary of Modeling Parameters allows you to check the modeling parameters just before generating the model.

   **Note**
   The screen Summary of Modeling Parameters contains an Advanced button. By clicking this button, you access the screen Advanced Model Parameters. For more information about these parameters, see the related topic.
• The **Model Name** is filled automatically. It corresponds to the name of the target variable (class for this scenario), followed by the underscore sign (_) and the name of the data source, minus its file extension (Census01 for this scenario).

• You have the possibility to display the results as a decision tree based on the five most contributive variables. To activate this option, check the box **Compute Decision Tree**.

• The **Autosave** button allows you to activate the feature that will automatically save the model once it has been generated. When the autosave option is activated, a green check mark is displayed on the **Autosave** button.

---

**Related Information**

[Setting the Advanced Parameters](#)

---

### 5.5.1.6.1 Activating the Autosave Option

The **Model Autosave** panel allows you to activate the option that will automatically save the model at the end of the generation process and to set the parameters needed when saving the model.

To activate the option, proceed as follows:

1. In the **Summary of Modeling Parameters** panel, click the **Autosave** button. The **Model Autosave** panel is displayed.
2. Check **Enable Model Autosave**.
3. Set the parameters listed in the following table.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Model Name</strong></td>
<td>This field allows you to associate a name with the model. This name will then appear in the list of models to be offered when you open an existing model.</td>
</tr>
<tr>
<td><strong>Description</strong></td>
<td>This field allows you to enter the information you want, such as the name of the training data set used, the polynomial degree or the performance indicators obtained. This information could be useful to you later for identifying your model. Note that this description will be used instead of the one entered in the <strong>Summary of Modeling Parameters</strong> panel.</td>
</tr>
</tbody>
</table>
| **Data Type** | This list allows you to select the type of storage in which you want to save your model. The following options are available:  
  ○ **Text Files**, to save the model in a text file.  
  ○ **Data Base**, to save the model in a database.  
  ○ **Flat Memory**, to save the model in the active memory. |
| **Folder** | Depending upon which option you selected, this field allows you to specify the ODBC source, the memory store or the folder in which you want to save the model. |
### 5.5.1.7 Setting the Advanced Parameters

On the screen *Summary of Modeling Parameters*, click the *Advanced* button. The screen *Advanced Model Parameters* appears.

#### 5.5.1.7.1 General Tab

The *General* tab allows you to define the general settings of the model, that is, the degree of the model, the score bin count, the number of correlations to display and the target key value.

#### 5.5.1.7.1.1 Defining the Degree of the Model (optional)

The model generated by Modeler - Regression/Classification is represented by a polynomial. This polynomial may be of degree 1, 2, 3 or greater. By defining the polynomial degree, you will define the degree of complexity of the model.

It is greatly recommended that you always use a degree of “1” (default value) for the first analysis of a dataset. Using a higher degree of polynomial does not guarantee that you will in all cases obtain a more powerful model.

For more information about the polynomial degree, see the related topic. For this scenario, keep the polynomial degree set to the default value – that is “1”.

To define the degree of the model, in the *Polynomial degree* field, enter the value corresponding to the degree of complexity of the model that you want to obtain.

**Related Information**

*Representation of a Model [page 36]*
5.5.1.7.1.2 Setting the Score Bin Count

This option allows you to define the number of bins to create for the score. This value must be set between 20 and 100 since a lower or higher number of bins would lead to poor model quality.

5.5.1.7.1.3 Exclusion of Low KR Variables

This option allows you to enable the exclusion of variables based on the value of their prediction confidence (KR). The application uses an internally computed threshold to decide whether a variable has a low prediction confidence. This threshold depends mainly on the data set size and target distribution.

To automatically exclude variables with low prediction confidence, check the option Exclusion of Low Prediction Confidence Variables.

5.5.1.7.1.4 Defining the Number of Correlations to Display

The section Correlations Settings allows you to set the parameters for the Correlation debriefing panel. That is, to select how many correlations should be displayed in that panel.

To say that variables are correlated implies that they each contribute some of the same information with respect to the target variable. A correlation contains two variables and a correlation rate. When you modify the number of correlations to display, the engine excluded the ones with the lowest correlation rate, thus keeping only the more significant ones.

5.5.1.7.1.5 Enabling the Post-Processing

This section allows setting some regression parameters according to three strategies. This option can only be activated when the model contains at least one continuous target variable.
The description of these strategies and an example of performance curve for each strategy are provided in the table below.

<table>
<thead>
<tr>
<th>Regression Strategy</th>
<th>Description</th>
<th>Example of Performance Curve</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without post-processing</td>
<td>The first strategy consists in disabling the regression post-processing during the learning model phase in order to create a regression similar to the one used in previous versions of the application. In this case, a standard regression is performed. No special improvement is made to the final scores. Original target values are used and raw score values are produced as outputs.</td>
<td><img src="image1" alt="Performance Curve for Without post-processing" /></td>
</tr>
<tr>
<td>With original target encoding</td>
<td>The second strategy, which applies to regressions using a post-processing, consists in using the original target value during the learning model phase to compute regression coefficients. The result of the regression is then transformed to align target segment means and score segment means in the post-processing phase. Note: This is the default strategy.</td>
<td><img src="image2" alt="Performance Curve for With original target encoding" /></td>
</tr>
</tbody>
</table>

*Note: These strategies are applicable to Classification, Segmentation, and Clustering Scenarios.*
<table>
<thead>
<tr>
<th>Regression Strategy</th>
<th>Description</th>
<th>Example of Performance Curve</th>
</tr>
</thead>
<tbody>
<tr>
<td>With uniform target encoding</td>
<td>The last strategy, which applies to regressions using a post-processing, consists in using first an encoded target value instead of the original target value during the learning model phase in order to have a uniform distribution: it is the pre-processing phase. Then, regression coefficients are computed and scores are transformed in the original target space during the post-processing phase.</td>
<td><img src="image" alt="Example of Performance Curve" /></td>
</tr>
</tbody>
</table>

**Note**
This strategy is to be preferred when the default strategy does not produce models with enough quality, which is often the case with very skewed target distributions.

### 5.5.1.7.1.5.1 Regression Without Post-processing

Uncheck the option *Enable Post-processing*.

- [x] Enable post-processing
- [ ] Original target encoding
  - [ ] Uniform target encoding

**Note**
It is not possible to change the target encoding strategy when the post-processing is disabled.

### 5.5.1.7.1.5.2 Regression with Original Target Values

1. Check the option *Enable Post-processing*.
2. Select the radio button *Original target encoding*.

- [x] Enable post-processing
- [ ] Original target encoding
  - [ ] Uniform target encoding
This strategy is set by default.

5.5.1.7.3.3 Regression with Uniform Target Encoding

1. Check the option *Enable Post-processing*.
2. Select the radio button *Uniform target encoding*.

5.5.1.7.6 Defining the Target Key Values

For the binary targets you have the option to select which value is the key category for each target. By default, the category selected by the application is the least represented in the dataset.

The *Advanced Model Parameters* screen lists all the binary targets of the current model allowing you to define the key category for each target, that is the expected value of the target.

In this scenario, do not define a value for the target variable. The application will automatically select “1” as the key category for the Class variable.

To define the key category value for a target variable, in the *Target Key* field corresponding to the chosen target, enter the key value.

5.5.1.7.2 Auto-selection Tab

The *Auto-selection* tab allows you to define the parameters of the automatic variable selection.

Setting the Variables Auto-Selection

The section *Auto-selection* allows you to automatically reduce the number of variables in the model in relation to quality criteria. This selection is done by successive iterations. There are two selection modes, one based on the number of variables to keep, and the other on the amount of information that should be kept. In this instance, the information is the sum of the variables contributions.
To use the auto-selection, check the box Enable Auto-Selection. The corresponding options are activated. By default the parameters are set to: Select the best model keeping between 1 and all variables. Any parameter that can be changed is marked as a hyperlink (blue, underlined).

**Choosing the Selection Mode**

To Select the Selection Mode

1. Click the link indicating the type of model to keep. For example, the best model in the sentence: Select the best model keeping between 1 and all variables.
   - A drop-down menu is displayed offering the following options:
     - the best model
     - the last model
2. Select the desired option.
3. Click OK.

Selecting the number of variables is mandatory and allows fixing the minimum and the maximum number of variables in the final model. To select the number of variables

1. In the sentence defining the number of variables Select the best model keeping between 1 and all variables, select the minimum number of variables (for example, 1 variable) and the maximum number of variables (for example, all variables).
2. For the minimum number of variables, a slide is displayed ranging from 1 to the total number of variables in the model.
   - Move the cursor on the slide to select the quantity of your choice.
   - For the maximum number of variables you can either confirm the minimum number selected previously by choosing Keep all variables, or choose a maximum number of variables.
3. Click OK.

**Choosing the Stopping Criteria**

Choose between two variable selection parameters:

- **Each step removes 1 variable**
  - This option allows you to set the number of variables that should be excluded at each iteration.

- **Each step keeps 95.0% of information**
  - This option allows you to set the amount of information that should be kept at each iteration (thus limiting the loss of information).

To select the number of variables:

1. Click the link indicating the number of variables in the sentence Each step removes 1 variable. A slide is displayed ranging from 1 to the total number of variables in the model.
2. Move the cursor on the slide to select the number of your choice.
3. Click OK.
To select the information amount:

- Click the link indicating the amount of information to keep in the sentence *Each step keeps 95.0% of information*. A slide is displayed.
- Move the cursor on the slide to select the quantity of your choice.
- Click OK.

The quality loss can be set in the sentence *Search process stops with a drop of 1.0% of KI and KR*. To set the authorized quality loss:

1. Click the link indicating the percentage of loss (for example, 5.0%). A slide is displayed.
2. Select the maximum percentage of authorized quality loss with the cursor.
3. Click OK.
4. Click the quality criterion. A drop-down list is displayed offering the following options:
   - Based on $KI + 2KR$, the quality loss is based on both the predictive power ($KI$) and twice the prediction confidence ($KR$).
   - $KI$ and the $KR$, the quality loss is limited for both the predictive power ($KI$) and the prediction confidence ($KR$). It is the default value.
   - $KI$, the quality loss is limited for the predictive power ($KI$) only.
   - $KR$, the quality loss is limited for the prediction confidence ($KR$) only.
5. Select the option of your choice.
6. Click OK.

### 5.5.1.7.3 Risk Mode Tab

This tab allows you to select a specific learning mode for your model.

#### To Enable the Risk Mode

1. Select the tab *Risk Mode*.
2. Check the box *Enable*. The tab activates and the *Risk Mode* settings are displayed.

### 5.5.1.7.3.1 Setting Risk Mode

Risk Mode allows advanced users to ask a classification model to translate its internal equation obtained with no constraints into a specified range of scores associated with good/bad odds ratio.

When this mode is activated, the different encodings that are used internally for continuous and ordinal variables are merged in a single representation, allowing a simpler view of the model internal equations. This is particularly useful when the usage of predictive model is subject to legal restrictions: the model equations are now simple enough to be understood by legal departments, and can be exposed, not only in programming language, as it was already the case before, but even in simple words.

The underlying technology is also used to display so called ‘score cards’.
To use this mode, you need to choose:

- a **Risk Score** associated with a **Good/Bad Odds ratio**

  **Note**
  
The odds ratio is the ratio between 'good' and 'bad', i.e. \((1-p)/p\) where \(p\) is the probability of risk.

- the number of **Points to double the odds**

  **Note**
  
The points to double the odd (PDO) are the number of risk points needed to double the odds ratio.

For example, considering a **Risk Score** equal to 615, an odds ratio of 9 : 1 and 15 points to double the odds. In this case, the application automatically re-scales the internal scores to scores in **Risk Mode** space and associates an odds ratio to each score in **Risk Mode** space.

In this scenario, do not activate the Risk mode.

**To Define the Risk Mode Parameters**

1. In the field **Risk Score**, enter the score you want to associate with a good/bad odds ratio.
2. In the field for **good/bad odds ratio of**, enter the ratio.
3. Indicate the increase of score points needed to double the odds in the field **Points to double odds**.
4. Click the button **View Score Table** to display the table of scores associated with the corresponding good/bad odds ratio.

![Image of View Score Table]

**5.5.17.3.2 Risk Fitting Domain**

This option allows the user to control the way risk score fitting is performed, that is, how InfinitelInsight® fits its own scores to the risk scores.

The risk fitting has two modes:

- **PDO based**: the area equals \([\text{Median Score} - N\times\text{PDO} ; \text{Median Score} + N\times\text{PDO}]\). \(N\) (number of PDOs around the median score) must be specified by the user. By default, \(N\) is set to 2.
PDO stands for Points to double the odds.

- **Frequency based**: the area equals \([\text{Quantile}(\text{Freq}) : \text{Quantile}(1.0 - \text{Freq})]\). The frequency of higher and lower scores to be skipped must be specified by the user. By default, the frequency is set to 15%.

If you do not check the box **Risk Fitting Domain**, the mode Frequency based will be used by default.

The fitting can be weighted or not.

### 5.5.1.7.3.2.1 To Set the Risk Fitting Parameters

1. Check the box **Risk Fitting Domain**.
2. Select the mode you want to use.
3. Depending on the selected mode, set the appropriate value in the corresponding field.
4. If you want to use weights for the fitting, check the box **Use Score Bin Frequency as Weights**.

### 5.5.1.7.4 Gain Chart Tab

This tab allows you to compute the gain chart on the training dataset, that is to rank your data in order of descending scores and split it into exact quantiles (decile, vingtile, percentile). This option can be useful to check the model performance on the validation dataset.

To Compute the Gain Chart

1. Select the tab **Gain Chart**.
2. Check the box Compute Gain Chart on Training Data.
3. In the list, select the Number of Quantiles you want your data to be segmented in.
4. You can add additional variables in order to estimate profits per segments of the population:
   1. In the Variables list, select the variables you want to add to the gain chart. Use the CTRL key to select multiple variables.
   2. Click the > button to add the selected variables to the list Values for Gain Chart.
5. The sum of each selected variable will be calculated for each segment of the population.
6. Click OK to save the advanced parameters and go back to the panel Summary of Modeling Parameters.

The result of the gain chart computation is available after the model has been generated in the Statistical Reports panel, section Model Performance.

5.5.2 Step 2 - Generating and Validating the Model

Once the modeling parameters are defined, you can generate the model. Then you must validate its performance using the predictive power (KI) and the prediction confidence (KR):

- If the model is sufficiently powerful, you can analyze the responses that it provides in relation to your business issue (see Step 3 - Analyzing and Understanding the Model Generated), and then apply it to new datasets (see Step 4 - Using the Model).
- Otherwise, you can modify the modeling parameters in such a way that they are better suited to your dataset and your business issue, and then generate new, more powerful models.

Related Information

Step 3 - Analyzing and Understanding the Model Generated [page 95]
Step 4 - Using the Model [page 131]

5.5.2.1 Generating the Model

To generate the model:

1. On the Advanced Model Parameters screen, click Generate.

   The screen Training the Model appears. The model is being generated. A progress bar allows you to follow the process.

2. If the Autosave option has been activated in the panel Summary of Modeling Parameters, a message is displayed at the end of the learning process confirming that the model has been saved.

3. Click Close.

4. Once the model has been generated, click Next to go to panel Using the Model.
5.5.2.2 Following the Progress of the Generation Process

There are two ways for you to follow the progress of the generation process:

- The Progress Bar displays the progression for each step of the process. It is the screen displayed by default.
- The Detailed Log displays the details of each step of the process.

To display the Progression Bar

Click View Type and select (Progress).
The progression bar screen appears.

To Display the Detailed Log

Click View Type and select (Log) button.
The detailed log displays the details of each step of the process.

To Stop the Learning Process

1. Click the (Stop Current Task) button.
2. Click the Previous button.
The screen Summary of Modeling Parameters appears.
3. Go back to the section on checking modeling parameters.

Related Information

Checking Modeling Parameters [page 82]

5.5.2.3 Validating the Model

Once the model has been generated, you must verify its validity by examining the performance indicators:

- The predictive power allows you to evaluate the explanatory power of the model, that is, its capacity to explain the target variable when applied to the training dataset. A perfect model would possess a predictive power equal to 1 and a completely random model would possess a predictive power equal to 0.
The prediction confidence defines the degree of robustness of the model, that is, its capacity to achieve the same explanatory power when applied to a new dataset. In other words, the degree of robustness corresponds to the predictive power of the model applied to an application dataset.

For this scenario, the model generated has the following performance indicators:

- A quality indicator $K_I$ equal to 0.808,
- A robustness indicator $K_R$ equal to 0.992.

The model performs sufficiently well. You do not need to generate another.

### To Validate the Model Generated

1. Verify the Predictive Power ($K_I$) and Prediction Confidence ($K_R$) of the model.
   - If the performance of the model meets your requirements, go to Step 3 - Analyzing and Understanding the Model Generated.
   - Otherwise, go to the procedure To Generate a New Model.
2. You can also check other indicators provided in addition to $K_I$ and $K_R$ during the model generation. For example, you could view the total elapsed time required to generate the model and information on the standard error rate.

### To Generate a New Model

You have two options. On the screen *Training the Model*, you can:

- Either click the *Previous* button to return to the modeling parameters defined initially. Then you can modify the parameters one by one.
- Or click the *Cancel* button to return to the main screen of the modeling assistant. Then you must redefine all the modeling parameters.

### Related Information

Step 3 - Analyzing and Understanding the Model Generated [page 95]

### 5.5.3 Step 3 - Analyzing and Understanding the Model Generated

The suite of plotting tools within the application allows you to analyze and understand the model generated:

- The performance of the model with respect to a hypothetical perfect model and a random type of model,
- The contribution of each of the explanatory variables with respect to the target variable,
• The significance of the various categories of each variable with respect to the target variable.

5.5.3.1 Presentation of the User Menu

Once the model has been generated, Click Next. The screen Using the Model appears.

The screen Using the Model presents the various options for using a model, that allow you to:

• Display the information relating to the model just generated or opened (Display section), referring to the model curve plots, contributions by variables, the various variables themselves, HTML statistical reports, table debriefing. Some information is only displayed upon request from the user: the display of Modeler - Regression/Classification results as a decision tree, which can be specified in the modeling parameters before the model generation, or the display of model parameters, which can be requested in the general user options.

• Apply the model just generated or opened to new data, to run simulations, and to refine the model by performing automatic selection of the explanatory variables to be taken into consideration (Run section).

• Save the model, or generate the source code (Save/Export section).

5.5.3.2 Model Overview

The Model Overview screen displays the same information as the training summary.

Overview

<table>
<thead>
<tr>
<th>Name</th>
<th>Name of the model, created by default from the target variable name and the dataset name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset</td>
<td>Name of the dataset</td>
</tr>
<tr>
<td>Initial Number Variables</td>
<td>Number of explanatory variables used</td>
</tr>
<tr>
<td>Number of Selected Variables</td>
<td>Number of explanatory variables actually used by the resulting model.</td>
</tr>
<tr>
<td>Number of Records</td>
<td>Number of records in the dataset</td>
</tr>
<tr>
<td>Building Date</td>
<td>Date and time when the model was built</td>
</tr>
<tr>
<td>Learning time</td>
<td>Total learning time</td>
</tr>
</tbody>
</table>
### Name

<table>
<thead>
<tr>
<th>Name</th>
<th>Name of the model, created by default from the target variable name and the dataset name</th>
</tr>
</thead>
</table>
| Engine name | Depending on the feature used:  
  - Kxen.RobustRegression  
  - Kxen.SmartSegmenter  
  - Kxen.TimeSeries  
  - Kxen.AssociationRules  
  - Kxen.EventLog  
  - Kxen.SequenceCoder  
  - Kxen.SocialNetwork |

---

### Modeling Warnings

<table>
<thead>
<tr>
<th><strong>Monotonic Variables Detected</strong></th>
<th>Indicates if monotonic variables have been found in the dataset, that is, variables which direction of variation is constant, in the reading order of the data in the estimation dataset.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Suspicious Variables Detected</strong></td>
<td>This report presents a list of variables that are considered to be suspicious. These suspicious variables have a predictive power over 0.9, they are very correlated to the target variable. This means these variables probably bring a biased information and should not be used for the modeling. A special attention should be taken towards those variables. A more detailed report lists which variables exactly are suspicious and at which extent (see <a href="#">Statistical Reports</a> <a href="#">Expert Debriefing</a> <a href="#">Suspicious Variables</a>).</td>
</tr>
</tbody>
</table>

---

### Targets

#### For each nominal variable

<table>
<thead>
<tr>
<th>&lt;Name&gt;</th>
<th>Name of the target variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target key</td>
<td>Wanted target value</td>
</tr>
<tr>
<td>Target categories Frequency</td>
<td>Percentage of all the target value in the Estimation dataset, when dealing with a nominal target</td>
</tr>
</tbody>
</table>

#### For each continuous target variable:

<table>
<thead>
<tr>
<th>&lt;TargetName&gt;</th>
<th>Name of the target variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>Minimum value found for the target variable in the Estimation dataset</td>
</tr>
<tr>
<td>Max</td>
<td>Maximum value found for the target variable in the Estimation dataset</td>
</tr>
<tr>
<td>Mean</td>
<td>Mean of the target variable values on the Estimation dataset</td>
</tr>
</tbody>
</table>
### Performance Indicators

For each target

<table>
<thead>
<tr>
<th>Target Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>\texttt{rr_&lt;TargetName&gt;}</td>
<td>Target name</td>
</tr>
<tr>
<td>\texttt{kc_&lt;TargetName&gt;}</td>
<td>Note that \texttt{rr} indicates a regression/classification and \texttt{kc} indicates a segmentation/clustering.</td>
</tr>
</tbody>
</table>

#### Predictive Power (KI)

Quality indicator that corresponds to the proportion of information contained in the target variable that the explanatory variables are able to explain.

#### Prediction Confidence (KR)

Robustness indicator that signifies the capacity of the model to achieve the same performance when it is applied to a new dataset exhibiting the same characteristics as the training dataset.

### 5.5.3.2.1 Model Overview Options

#### To Copy the Model Overview

Click the \(\text{(Copy)}\) button.

The application copies the HTML code of the screen. You can paste into a word processing or spreadsheet program, a text editor, ...

#### To Save the Model Overview

Click the \(\text{(Save)}\) button situated under the title.

The file is saved in HTML format.

#### To Print the Model Overview

1. Click the \(\text{(Print)}\) button situated under the title.
A dialog box appears, allowing you to select the printer to use.

2. Select the printer to use and set other print properties if need be.
3. Click OK.
   The report is printed.

**To Export to PowerPoint**

Click the (Export to PowerPoint) button.

**5.5.3.3 Model Graphs**

Depending on the type of the target, the model graph plot allows you to:

- View the realizable profit that pertains to your business issue using the model generated when the target is nominal.
- Compare the performance of the model generated with that of a random type model and that of a hypothetical perfect model when the target is nominal.
- Compare the predicted value to the actual value when the target is continuous.

On the plot, for each type of model, the curves represent:

- When the target is nominal, the realizable profit (on the Y axis) as a function of the ratio of the observations correctly selected as targets relative to the entire initial dataset (on the X axis).
- When the target is continuous, the predicted value or score (on the X axis) in respect with the actual value or target (on the Y axis).

**To Display the Model Graph**

1. On the screen *Using the Model*, click the *Model Graphs* option.
The model graphs appear. When the target is nominal, the following curve is displayed:

The default parameters display the profit curves corresponding to the Validation sub-set (blue line), the hypothetical perfect model (Wizard, green line) and a random model (Random, red line). The default setting for the type of profit parameter is Detected profit, and the values of the abscissa are provided in the form of a percentage of the entire dataset.

When the target is continuous, the following curve is displayed:

The default parameters display the curves corresponding to the Validation sub-set (blue line) and the hypothetical perfect model (Wizard, green line). The blue area represents the standard deviation of the current model. For more information on the meaning of model curves, see the topic on understanding model graphs.

2. When there is more than one target, select the target for which you want to see the curves in the Models list.
Note
To each variable corresponds a model. The name of each model is built from the \textit{rr} (Robust Regression) prefix and the model target name.

3. Select the viewing options that interest you.

Related Information

Understanding Graphs for a Model with a Nominal Target [page 104]
Understanding Graphs for a Model with a Continuous Target [page 105]

5.5.3.3.1 Plot Options

To Display the Graphs for the Estimation, Validation and Test Sub-sets

Click \textit{Dataset} and select one of the following options that allow you to switch between:

- the graph for the Validation sub-set:
- the graphs for all the sub-sets:

To Change the View Type

Click \textit{View Type} and select the desired option.

To Copy the Model Graph

Click the (Copy) button and select the desired option.

The application copies the parameters of the plot. You can paste it into a spreadsheet program (such as Excel) and use it to generate a graph.

To Save the Model Graph

1. Click the (Save) button.
A dialog box appears, allowing you to select the file properties.
2. Type a name for your file.
3. Select the destination folder.
4. Click OK.
   The plot is saved as a PNG formatted image.

To Print the Model Graph

1. Click the (Print) button situated under the title.
   A dialog box appears, allowing you to select the printer to use.
2. Select the printer to use and set other print properties if need be.
3. Click OK.
   The report is printed.

To Export the Model Graph to Microsoft Excel

Click the (Export to Excel) button situated under the title. An Excel sheet opens containing the model graph you are currently viewing along with its data.

To Open the Current Graph in a New Window

Click the (Pin View) button. The current graph is displayed in a new window.

5.5.3.3.1.1 Plot Options for a Nominal Target

On the model curve plot, different options allow you to visualize:
- Exact profit values for a point for all the displayed curves.
- The curves for the different profit types: Detected, Lift, Normalized, and Customized.
For more information on profit types, see the related topic.

To Display the Exact Profit Values for a Given Point

On the screen Model Curves, on the plot, click a point on one of the curves presented. For instance, by clicking a point on any one of the curves whose value on the abscissa is 25%, the exact profit values appear.
To Select a Profit Type

1. On the screen Model Curves, beneath the plot, click the drop-down list associated with the Profit field. The list of profit types appears.
2. Select a profit type.
   The corresponding profit curves appear.

Related Information

Profit Type [page 46]

5.5.3.3.1.2 Plot Options for a Continuous Target

To Display the Exact Profit Values for a Given Point

On the screen Model Graphs, on the plot, click a point on one of the curves presented.

To Select the Debriefing Type

1. On the screen Model Graphs, above the plot, click the drop-down list associated with the Debriefing Type field.
   The list of debriefing types appears.
2. Select a debriefing type.
   The corresponding plot appears.
5.5.3.3.2 Understanding Model Graphs

5.5.3.3.2.1 Understanding Graphs for a Model with a Nominal Target

The following figure represents the model graph produced using the default parameters.

On the plot, the curves for each type of model represent the profit that may be realized (Y axis), that is, the percentage of observations that belong to the target variable, in relation to the number of observations selected from the entire initial dataset (X axis). On the X axis, the observations are sorted in terms of decreasing "score", that is, the decreasing probability that they belong to the target category of the target variable.

In the application scenario, the model curves represent the ratio of prospects likely to respond in a positive manner to your marketing campaign, relative to the entire set of prospects contained in your database.

Detected profit is the default setting for type of profit. Using this type of profit:

- The value "0" is assigned to observations that do not belong to the target category of the target variable,
- The value "1/(frequency of the target variable in the dataset)" is assigned to observations that do belong to the target category of the target variable.

The following table describes the three curves represented on the plot created using the default parameters.
### Understanding Graphs for a Model with a Continuous Target

The following graph represents the model curve plot produced using a continuous target.

![Model Curve Plot](image)

### The curve... Represents... For instance, by selecting...

<table>
<thead>
<tr>
<th>Wizard (green curve, at the top)</th>
<th>The profit that may be achieved using the hypothetical perfect model that allows one to know with absolute confidence the value of the target variable for each observation of the dataset</th>
<th>25% of the observations from your entire dataset with the help of a perfect model, 100% of observations belonging to the target category of the target variable are selected. Thus maximum profit is achieved. Note that these 25% correspond to the proportion of prospects who responded in a positive manner to your marketing campaign, during your test phase. For these prospects, the value of the target variable, or profit, is equal to 1.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validation (blue curve, in the middle)</td>
<td>The profit that may be achieved using the model generated by Modeler - Regression/Classification that allows one to perform the best possible prediction of the value of the target variable for each observation of the dataset</td>
<td>25% of the observations from your initial dataset with the help of the model generated, 66.9% of the observations belonging to the target category of the target variable are selected.</td>
</tr>
<tr>
<td>Random (red curve, at the bottom)</td>
<td>The profit that may be achieved using a random model that does not allow one to know even a single value of the target variable for each observation of the dataset</td>
<td>25% of the initial dataset using a random model, 25% belonging to the target category of the target variable are selected.</td>
</tr>
</tbody>
</table>
The default graphic displays the actual target values as a function of predicted target values. Two curves are displayed: one for the Validation sub-set (blue line) and another for the hypothetical perfect model (Wizard, green line). The Validation curve gives Actual Target value as a function of Predicted Target value. For example, when the model predicts 35, the average actual value is 37. The Wizard curve is just X=Y, meaning that all the predicted values are equal to the actual values. The graph is an easy way to quickly see model error. When the curve is going far from Wizard, it means that the predicted value is suspicious.

The graph is computed as follow:

- about 20 segments or bins of predicted values are built. Each of these segments represents roughly 5% of the population.
- for each of these segments, some basic statistics are computed on actual value, such as the mean of the segment (SegmentMean), the mean of the associated target (TargetMean) and the variance of this target within that segment (TargetVariance). For example for predicted value in [17; 19], the mean would be 18.5, the actual target mean would be 20.5 and the actual target variance would be 9. In this case we could say that, if the predicted value is between 17 and 19, the model is underestimating a bit the actual value.

For each curve, a dot on the graph corresponds to the segment mean on the X-axis, and the target mean on the Y-axis.

The blue area represents the expected deviation of the current model. The blue area shows where about 70% of the actual values are expected to be. In other words, it means that, in case of a Gaussian distribution, about 70% of the actual points should be in the blue area (keep in mind that this is a theoretical percentage that may not be observed every time). The default setting for the type of curve parameter is Predicted versus Actual. The extreme values for prediction ranges are \( \text{TargetMean} - (\sqrt{\text{TargetVariance}}) \) ; \( \text{TargetMean} + (\sqrt{\text{TargetVariance}}) \).

\[ \text{Note} \]

\( \sqrt{\text{TargetVariance}} \) is equal to the Standard Deviation.

### 5.5.3.3.3 Predictive Power, Prediction Confidence and Model Graphs

On the model graph plot:

- Of the estimation data set (default plot), the predictive power corresponds to “the area found between the curve of the model generated and that of the random model” divided by “the area found between the curve of the perfect model and that of the random model”. As the curve of the generated model approaches the curve of the perfect model, the value of the predictive power approaches 1.
- Of the estimation, validation and test data sets (select the corresponding option from the list Data set, located below the plot), the prediction confidence corresponds to one minus “the area found between the curve of the estimation data set and that of the validation data set” divided by “the area found between the curve of the perfect model and that of the random model”.

5.5.3.4 Contributions by Variables

The Contributions by Variables plot allows you to examine the relative significance of each of the variables within the model. On this plot, each bar represents the contribution of an explanatory variable with respect to the target variable.

The following four types of plots allow you to visualize contributions by variables:

- **Variable Contributions**, that is, relative importance of each variable in the built model.
- **Variable Weights**, that is, weights (in the final polynomial) of the normalized variables.
- **Smart Variable Contributions**, that is, the variables internal contributions.
- **Maximum Smart Variable Contributions**, that is, the maximum smart variable contributions including only the maximum of similar variables. For example, only binned encoding of the continuous variable age will be displayed. This is the chart displayed by default.

Displaying Contributions by Variables

1. On the screen Using the Model, click the option Contributions by Variables.
   The plot Contributions by Variables appears. The default plot type is Maximum Smart Variable Contributions.
   If your dataset contains date or datetime variables, automatically generated variables can appear in this panel. For more information, refer to the section on date and date-time variables in the topic on variable storage formats.
2. You can drill down on a variable, that is, display the plot of details of this variable, where the categories of the variable can be seen. To zoom in on a variable, double-click the corresponding bar. Go to section **Significance of Categories**.

**Understanding Contributions by Variables**

Only the plot, *Maximum Smart Contributions by Variables*, the default selection, is presented in this guide.

The *Contributions by Variables* option allows the user to examine the relative significance of each of the explanatory variables in relation to the target variable. This significance is relative, as the weight of each variable is pro-rated as a function of the significance of the other explanatory variables.

The plot above corresponds to the model generated, and illustrates the two variables that contribute the most to the target variable, which, in this scenario, are:

- marital-status,
- capital-gain.

In other words, the marital-status and capital-gain variables are those which have the greatest effect on whether a prospect will respond positively or negatively to your marketing campaign. Among all the variables...
included in the sample dataset, these two are the most discriminatory variables with respect to the target variable *Class*.

**Correlated Variables**

To say that variables are correlated implies a certain level of redundancy, that they each contribute some of the same information with respect to the target variable. Two variables said to be highly correlated would describe the same information, or the same concept, to an even greater degree.

The plot *Smart Variable Contributions* reflects the correlation that may exist between various explanatory variables. When two variables A and B are strongly correlated:

- Variable A, with a greater contribution than B with respect to the target variable, becomes the "primary variable": the plot displays all its information, including what it has in common with variable B.
- Variable B, with a smaller contribution than A with respect to the target variable, becomes the "secondary variable": only its marginal contribution is displayed on the plot, meaning that only the supplementary contribution to target variable information, or the values that B does not share with A, are displayed. This difference of information is noted \[variable_B]-[variable_A].

**Encoded Variables**

Creating a model uses not only the original variables but also, in case of continuous or ordinal variables, their value as encoded by the application. This is called dual-encoding and allows the application to find all the information contained in each variable.

The encoded variables appear on the variable contributions plots with the prefix c_. For example, the encoded version of a continuous variable named age is noted c_age.

**i Note**

In Modeler, on the *Data Description* panel, if you enable the *Natural Encoding* for a given variable, its value encoded by the data encoding engine (c_variableName) will not be generated.

**Related Information**

*Variable Storage Formats [page 30]*
5.5.3.5 Category Significance

Definition

The Significance of Categories plot illustrates the relative significance of the different categories of a given variable with respect to the target variable.

Displaying the Significance of Categories Plot

1. On the screen Using the Model, click Category Significance. The plot Category Significance appears.

2. In the Variables list located above the plot, select the variable for which you want to display the categories. If your dataset contains date or datetime variables, automatically generated variables can appear in the Variables list. For more information, refer to section on date and date-time variables in the topic on variable storage formats.

Note

- You can display the relative significance of the categories of a variable directly from the plot Contributions by Variables. On the plot Contributions by Variables, double-click the bar of the variable which interests you.
In case no user structure has been defined for a continuous variable, the plot category significance displays the categories created automatically using the band count parameter. The number of categories displayed corresponds to the value of the band count parameter. For more information about configuring this parameter, please refer to the section Band Count for Continuous Variables.

Plot Options

To Switch Between "Validation Data Set" and "All Data Sets" Plots

1. Click Data Sets and select the (All Data Sets) button to display all datasets. The plot displaying all datasets appears.
2. Click Data Sets and select the (Validation Only) button to go back to the Validation Data Set plot.

To Switch between Curve and Bar Charts

1. Click View Type and select the button to display the curve chart. The curve plot appears.
2. Click View Type and select the button to go back to the bar chart.

Note

You can combine the different types of plot. For example, you can display All Datasets in a curve chart or the Validation Data Set in a bar chart.

Understanding the Plots of Variables

For this scenario, select the variable marital-status, which is the explanatory variable that contributes the most to the target variable Class. This plot presents the effect of the categories of the marital-status variable on the target variable. For an explanation of the plot, see the topic on variable categories and profit.

Related Information

Variable Storage Formats [page 30]
Band Count for Continuous Variables [page 75]
Variable Categories and Profit [page 112]
5.5.3.5.1 Variable Categories and Profit

The *Category Significance* plot shows the relative significance of the different categories of a given variable with respect to the target variable.

Each bar on the plot shows the amount of influence that category has on the target category (or hoped-for value) of the target variable. The bars are ordered by their amount of influence. In this example, the *Bar Orientation* option shows the bars vertically, so the bar the furthest to the left represents the category with the greatest positive effect, and the bar the furthest to the right, the category with the least effect.

If the *Bar Orientation* option is set to show the bars horizontally, the highest bar on the plot represents the category with the greatest positive effect. In other words, the higher a category appears on the plot, the more representative that category is of the target category of the target variable.
The length and direction of a bar correspond to the profit contributed by that category, in other words, whether the category has more or fewer observations belonging to the target category. A positive bar (Influence on Target greater than 0.0) indicates that the category contains more observations belonging to the target category than the mean (calculated on the entire dataset). A negative bar (Influence on Target less than 0.0) indicates that the category contains fewer observations belonging to the target category than the mean.

Note

You can display the profit curve for the selected variable by clicking the (Display Profit Curve) button located in the tool bar under the title.

The importance of a category depends on both its difference to the target category mean and the number of represented cases. High importance can result from any of the following:

- A high discrepancy between the category and the mean of the target category of the target variable
- A minor discrepancy combined with a large number of records in the category
- A combination of both

Use the Variables pull down menu to select and graph any of the variables in the model. Use the tool bar located under the title to copy the coordinates to the clipboard, print the plot, or save it in PNG format. The values are normalized and their sum always equals to "0". Depending on the chosen profit strategy, or on the continuous target variables value type, you can obtain all positive importance or negative and positive importance.
5.5.3.5.1.1 Axes

The X-axis shows the influence of the variable categories on the target. The significance of the different numbers on the X-axis are detailed in the following table:

<table>
<thead>
<tr>
<th>Number on the X-axis</th>
<th>Indicates that the category has</th>
</tr>
</thead>
<tbody>
<tr>
<td>positive number</td>
<td>a positive influence on the target</td>
</tr>
<tr>
<td>0</td>
<td>no influence on the target (the behavior is the same as the average behavior of the whole population)</td>
</tr>
<tr>
<td>negative number</td>
<td>a negative influence on the target</td>
</tr>
</tbody>
</table>

The Y-axis displays the variable categories. Categories sharing the same effect on the target variable are grouped. They appear as follow [Category_a;Category_b;Category_c]. Categories not containing sufficient numbers to provide robust information are grouped in the KxOther category. When a variable is associated with too many missing values, the missing values are grouped in the KxMissing category. Both categories are created automatically by the application.

5.5.3.5.1.2 Category Importance Definition

The following definition applies to continuous targets; some wording may be simplified for binary targets. The formulas presented below can also be applied to the binary target case (use categories instead of segments in this case).

We consider the case where a Regression/Classification regression model is trained on a continuous target/ signal S with the help of an input variable X.

Regression/Classification starts by binning the continuous target S into B segments: S1,...,SB and by computing the basic statistics of the inputs and the cross statistics of the inputs with respect to the target.

We will suppose that the input X is a nominal (categorical) variable, though the whole process can be extended easily to the case of ordinal and continuous inputs.

We will suppose that X has N categories: X1, ..., XN. We are interested in assessing the importance of a category Xi with respect to the target S. The importance of a category depends on two factors:

- The fact that the distribution of the target for this category is significantly skewed towards high values or low values when compared with the distribution of the target on the entire population.
- The frequency of this category.

High importance can result from either of the following:

- A high discrepancy between the target distribution for cases associated to this category and the distribution of the target variable for the entire population.
- A minor discrepancy combined with a large number of records in the category
- a combination of both

The application uses a non-parametric setting in which the category importance is defined as:

\[
\text{CategoryImportance}(X_i) = \text{NormalProfit}(X_i) \cdot \frac{\text{Freq}(X_i)}{2}
\]
where:

- \( \text{normalProfit}(X_i) \) is the normal profit of category \( X_i \), see below for a definition
- \( \text{Freq}(X_i) \) is the global frequency of the category \( X_i \)
- \( Z \) is a normalization constant.

We give below the details of the computation of these quantities.

**Normal Profit**

Each category of the target \( S_j \) is associated with a profit \( \text{profit}(S_j) \) defined such that:

\[
\sum_{j=1}^{j=B} \text{profit}(S_j) \text{Freq}(S_j) = 0
\]

The profit of a target category is a value in the range \([-1; +1]\). It is defined the following way from the (cumulated) target category frequencies:

\[
\text{profit}(S_j) = 2 \sum_{k=1}^{j-1} \text{Freq}(S_k) + \text{Freq}(S_j) - 1
\]

The normal profit of a category \( X_i \) is then defined as:

\[
\text{NormalProfit}(X_i) = \sum_{j=1}^{j=B} \text{Profit}(S_j) \text{Prob}[S_j | X_i]
\]

Where \( \text{Prob}[S_j | X_i] \) is the conditional probability of observing the target category \( S_j \) in the variable category \( X_i \) (cross statistics):

\[
\text{Prob}[S_j | X_i] = \frac{\text{Freq}(S_j; X_i)}{\text{Freq}(X_i)}
\]

The fact that these formulas rely only on frequencies makes them resistant to any monotonic transformation of the target \( S \).

**Normalization Constant**

The normalization can be approximated for non pathological continuous targets (that is continuous targets without distribution peak (Dirac)) as:

\[
Z = P[S > \text{median}(S)] + (1 - P[S > \text{median}(S)])
\]

In most cases, a good approximation is \( = 0.25 \).
Normal Profit Properties

There are several interesting things to note about normal profit:

- The normal profit of category is independent of the target values themselves (user can change the target value through monotonic transformations; the normal profit of the categories with respect to this target will not change). This belongs to non-parametric metrics.
- A consequence of 1 is that this metric is resistant to outliers: when there are a few occurrences of the target with very high values with respect to the rest of the target value distributions, the notion of normal profit is not impacted.
- The weighted sum of the normal profit for all categories of a given variables will always be 0.

5.5.3.5.2 Grouping Categories

On the plot of details of a variable, categories may appear grouped.

When the option Optimal Grouping is enabled, the application groups those categories sharing the same effect on the target variable. In the plot below (example in English), for the variable education, the categories Doctorate and Prof-School are grouped. If the explanatory variable is continuous, the application identifies the points where behavioral changes occur with respect to the target variable and automatically crops the variable into intervals exhibiting homogeneous behavior with respect to the target.

For more information, please see the related topic.
When categories do not contain sufficient numbers to provide robust information, they are grouped in the KxOther category, that is created automatically.

When a variable is associated with too many missing values, the missing values are grouped in the KxMissing category, that is also created automatically.

To understand the value of the categories KxOther and KxMissing, consider the following example. The database of corporate customers of a business contains the variable "web address". This variable contains the Web site address of the corporate customers contained in the database. Some companies have a Web site; others do not. In addition, each Web site address is unique. In this case, the application automatically transforms the "web address" variable into a binary variable with two possible values: KxOther (the firm has a Web site) and KxMissing (the firm does not have a Web site).

**Related Information**

Optimal Grouping for All Variables [page 76]
5.5.3.6 Statistical Reports

Statistical Reports provide you with a set of tables that allow you a more detailed debriefing of your model. These reports are grouped in different levels of debriefing:

- The Descriptive Statistics, which provides the statistics on the variables, their categories and the datasets, as well as the variables cross-statistics with the target.

  **Note**
  - If your dataset contains date or datetime variables, automatically generated variables will appear in the statistical reports.
  - In the section *Cross Statistics with the Target(s)*, the number of displayed categories corresponds to:
    - The number of categories as defined in the user structure
    - The band count if no user structure has been defined
    For more information, see the related topic on bad count.
  - In the section *Grouped Cross Statistics with the Target(s)*, if the option *Optimal Grouping* is enabled, the number of displayed categories is lower than that defined:
    - In the user structure
    - By the parameter band count if no user structure has been defined.

- The Model Performance, in which you will find the model performance indicators, the variables contributions and the score detailed statistics.

- The Control for Deviations, which allows you to check the deviations for each variable and each variable category between the validation and test datasets.

- The Expert Debriefing, in which you will find more specialized performance indicators, as well as the variables encoding, the excluded variables during model generation and the reason for exclusion, and so on.

Related Information

Variable Storage Formats [page 30]
Band Count for Continuous Variables [page 75]

5.5.3.6.1 Variable Exclusion Cause

Statistical reports include the section *Variable Exclusion Causes*. For regression and classification models, this section presents the reason why any variable was excluded from the model.

- *Overall Exclusions* shows the variables excluded from the whole model.
- *Target Specific Exclusions* shows the variables excluded towards a particular target.

The table below shows the possible variable exclusion causes:
<table>
<thead>
<tr>
<th>Overall Exclusions / Target Specific Exclusions</th>
<th>Name</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Exclusions</td>
<td>Constant</td>
<td>The variable has only one value (continuous variables) or one category (nominal or ordinal variables) in the dataset. The variable is discarded with respect to all targets.</td>
</tr>
<tr>
<td>Overall Exclusions</td>
<td>Small variance</td>
<td>For continuous variables, the variance is small. The variable variation is noise. The variable is discarded with respect to all targets.</td>
</tr>
<tr>
<td>Target Specific Exclusions</td>
<td>Fully Compressed</td>
<td>The variable has been fully compressed with respect to the target. It will be excluded from the model with respect to this target.</td>
</tr>
<tr>
<td>Target Specific Exclusions</td>
<td>Small KI on estimation</td>
<td>The variable has a small KI on Estimation dataset with respect to the target. It will be excluded from the model with respect to this target.</td>
</tr>
<tr>
<td>Target Specific Exclusions</td>
<td>Small KI on validation</td>
<td>The variable has a small KI on Validation dataset with respect to the target. It will be excluded from the model with respect to this target.</td>
</tr>
<tr>
<td>Target Specific Exclusions</td>
<td>Large KI difference</td>
<td>A large KI difference has been observed for this variable between Estimation and Validation datasets with respect to the target. It will be excluded from the model with respect to this target.</td>
</tr>
<tr>
<td>Target Specific Exclusions</td>
<td>Small KR</td>
<td>The variable has a small KR with respect to the target. It will be excluded from the model with respect to this target.</td>
</tr>
</tbody>
</table>

### 5.5.3.6.2 Statistical Report Options

A toolbar is provided allowing you to modify how the current report is displayed, to copy the report, to print it, to save it or to export it to Excel.
## Display Options

<table>
<thead>
<tr>
<th>Menu</th>
<th>Option</th>
</tr>
</thead>
<tbody>
<tr>
<td>View</td>
<td>Display the current report view in the graphical table that can be sorted by column.</td>
</tr>
<tr>
<td></td>
<td>Display the current report view as an HTML table.</td>
</tr>
<tr>
<td></td>
<td>Some reports can be displayed as a bar chart. This bar chart can be sorted by ascending or descending values, or by ascending or descending alphabetical order. You can also select which data should be displayed.</td>
</tr>
<tr>
<td></td>
<td>Display certain reports as a pie chart.</td>
</tr>
<tr>
<td></td>
<td>Display certain reports as a line chart.</td>
</tr>
<tr>
<td>Sort</td>
<td>When the current report is displayed as a bar chart, this option allows you to change the orientation of the bars (from horizontal to vertical and vice versa).</td>
</tr>
<tr>
<td></td>
<td>Display the current report with no sorting.</td>
</tr>
<tr>
<td></td>
<td>Sort the current report by ascending values.</td>
</tr>
<tr>
<td></td>
<td>Sort the current report by descending values.</td>
</tr>
<tr>
<td></td>
<td>Sort the current report by ascending names.</td>
</tr>
<tr>
<td></td>
<td>Sort the current report by descending names.</td>
</tr>
<tr>
<td>Series</td>
<td>Select which columns to display for current report.</td>
</tr>
</tbody>
</table>

## Usage Options

<table>
<thead>
<tr>
<th>Menu</th>
<th>Option</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Copy the data from the current view of the displayed report. The data can then be pasted in a text editor, a spreadsheet, a word processing software.</td>
</tr>
<tr>
<td></td>
<td>Print the current view of the selected report depending on the chosen display mode (HTML table, graph, ...).</td>
</tr>
</tbody>
</table>
### Menu

<table>
<thead>
<tr>
<th>Option</th>
</tr>
</thead>
</table>
| ![Save under different formats (text, html, pdf, rtf) the data from the current view of the selected report.](image)
| ![Save under different formats (text, html, pdf, rtf) the data from all the views of the selected report.](image)
| ![Export to Excel.](image)
| ![Save all reports.](image)
| ![Save the customized style sheet.](image) |

#### 5.5.3.7 Scorecard

This screen provides you with the coefficients associated to each category for all variables in a regression model.

![Scorecard](image)
To Obtain a Score

Add all the coefficients corresponding to the selected value of each variable.

**Note**

In the case of a continuous variable, the scorecard always includes a number of categories that is higher than in the user defined structure or as given by the parameter *band count* if no user structure has been set. Indeed, the encoding of variables for the scorecard adds target curve points to increase the accuracy of coding according to the training dataset. These points split some existing categories and thus increase the number of categories in the scorecard.

### 5.5.3.7 Risk Mode

The representation of a model equation is easier to read and to interpret in the Risk Mode due to stepwise encoding for ordinal and continuous variables.

In the Risk Mode, it is easy to define which category has a negative or positive effect on the risk score and consequently on the odds or on the probability of risk.

In order to illustrate the advantages of a scorecard in interpreting results, the variable age will be used for this example.

The segment \([24;27]\) has a risk score of about 30 and the segment \([37;43]\) has a risk score of about 15. According to the parameter PDO (set in this example to 15), it is easy to conclude that the segment \([37;43]\) is two times more risky or that the odds of the segment \([37;43]\) are two times inferior to the segment \([24;27]\).
5.5.3.8 Confusion Matrix

You can use the Confusion Matrix to compare the predicted value of the target variable with its actual value.

When generating the model (see step Generating the Model [page 93]), you have set up the decision threshold and decided on the score above which the observations are considered as positive. A positive observation is an observation that belongs to the population you want to target. A negative observation is an observation that does not belong to this target population.
Understanding the Confusion Matrix

The threshold

There are three ways to set the threshold using the displayed slide bar:

- By selecting the percentage of population to target if the population is sorted by descending order of score (% of Population).
- By selecting the percentage of positive observations you want to detect (% of Detected Target).
- By selecting the score used to differentiate positive observations from negative ones (Score Threshold). Any observation with a score above the threshold is considered positive, on the contrary any observation with a score below the threshold is considered negative.

The slide is graduated from the lowest score (on the left) to the highest score (on the right). The values corresponding to each option are displayed under the slide.

When you move the cursor, the confusion matrix is updated accordingly.

The Confusion Matrix

The following table details how to read the confusion matrix.
Predicted [Target Category] Positive Observations Predicted
Predicted [Non-target Category] Negative Observations Predicted

True [Target Category] Actual Positive Observations
Number of correctly predicted positive observations (True Positive = TP)
Number of actual positive observations that have been predicted negative (False Negative = FN)

False [Non-target Category] Actual Negative Observations
Number of actual negative observations that have been predicted positive (False Positive = FP)
Number of correctly predicted negative observations (True Negative = TN)

By default, the Total Population is the number of records in the Validation dataset. You can modify this number to see the confusion matrix for the population on which you want to apply your model.

Example
Consider a classification system that organizes the e-mails into two categories: relevant e-mails and junk e-mails. You want to know how many relevant e-mails are wrongly identified as junk e-mails and how many junk e-mails are not identified as junk e-mails.

For the exercise, suppose that you are testing the classification system with 100 relevant e-mails and 100 junk e-mails.

The confusion matrix could be:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>True [Normal e-mails] Actual Positive Observations</td>
<td>95</td>
<td>5</td>
</tr>
<tr>
<td>False [Junk e-mails] Actual Negative Observations</td>
<td>3</td>
<td>97</td>
</tr>
</tbody>
</table>

And is read as follows:

- Among the 100 relevant e-mails, 95 e-mails are predicted as relevant e-mails and 5 are predicted as junk e-mails.
- Among the 100 junk e-mails, 3 are predicted as relevant e-mails and 97 are predicted as junk e-mails.
- Among the e-mails predicted as relevant e-mails (TP + FP), 3 are actually junk e-mails.
- Among the e-mails predicted as junk e-mails (FN + TN), 5 are actually relevant e-mails.

The Metrics:
The following table details how to read the confusion matrix.

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Definitions</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification Rate</td>
<td>Percentage of observations accurately classified by the model when applied on the training dataset.</td>
<td></td>
</tr>
</tbody>
</table>

Classification, Regression, Segmentation and Clustering Scenarios
Modeler - Regression/Classification
### Sensitivity
Percentage of actual positive observations that have been correctly predicted.

\[ TP / (TP + FN) \]

### Specificity
Percentage of negative observations that have been correctly predicted.

\[ TN / (FP + TN) \]

### Precision
Percentage of detected positive observations that are actually positive observations.

\[ TP / (TP + FP) \]

### Recall
Percentage of actual positive observations that have been detected as positive.

\[ TP / (TP + FN) \]

### F1 score
Harmonic mean of Precision and Recall (Recall and Precision are evenly weighted).

\[ F1 = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \]

---

**The Cost Matrix**

This section allows you to visualize your profit depending on the selected score, or to automatically select the score depending on your profit parameters.

For each observation category, enter a profit or a cost per observation. The total profit is automatically displayed on the right of the table.

To know the threshold that will give you a maximum profit for the profit parameters you have set, click the button **Maximize Profit**.

---

#### Example

In the following profit/cost table, each positive observation correctly identified will yield $15, but each negative observation identified as positive will cost you $8.

<table>
<thead>
<tr>
<th>Category</th>
<th>Predicted[1]</th>
<th>Predicted[0]</th>
</tr>
</thead>
<tbody>
<tr>
<td>True[1]</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>True[0]</td>
<td>-8</td>
<td>0</td>
</tr>
</tbody>
</table>

---

### 5.5.3.9 Decision Tree

The panel Decision Tree allows you to display the results of a regression or classification model generated by Modeler as a decision tree based on the five most contributive variables. You can only access the Decision Tree panel if you selected the option Compute Decision Tree before generating the model.

---

**To Display the Decision Tree for a Target**

In the Target list, select the target for which you want to display the decision tree.
Understanding the Decision Tree Panel

The panel Decision Tree is split into three parts:

1. The decision tree itself, which is displayed in the upper section of the panel
2. Two tabs located in the left bottom part of the panel provide you with information on the nodes and with the profit curve corresponding to the current decision tree
3. A navigator allowing you to visualize what part of the tree you are studying is displayed in the right bottom part of the panel.

The Decision Tree

Each node in the tree displays:

- The name of the expanded variable, for example Marital-status.
- The categories on which the node population has been filtered, for example {Married-AF-spouse;Never-married}
- The Population of the node
- The ratio of Positive Target (for nominal targets) or the Target Mean (for continuous targets)

When you go over a node, several options are offered:
The thickness of the arrows depends on the amount of population in the node. In the following example, the arrow leading to the node corresponding to the category $[0;4386]$ of capital-gain is thicker since the node population is significantly higher than the one from the node capital-gain $]4386;41310[$.

### Node Details

When you select a node, the node information is displayed in the tab **Node Details** located in the lower part of the panel. This tab indicates the target for which the current decision tree is displayed and provides you with the following information for each dataset in the model:

- **Population Count**, that is, the number of records found in the current node
- For continuous targets:
  - **Target Mean**, that is, the mean of the target for the current node.
- For nominal targets:
  - **Positive Target Count**, that is, the number of records for which the target is positive,
  - **Positive Target Ratio**, that is, the percentage of the node population for which the target is positive,
  - **Negative Target Count**, that is, the number of records for which the target is negative,
  - **Negative Target Ratio**, that is, the percentage of the node population for which the target is negative,
  - **Variance**, that is, the variance for the current node,
  - **Weighted Population**, that is, the number of records when using a weight variable.

### Profit Curve

The profit curve for the current decision tree is displayed in the tab **Profit Curve** located in the lower part of the panel. This profit curve changes with every modification made on the decision tree.
The profit curve corresponding to the node containing the whole population is equal to the random curve.

When you expand the node with the highest percentage of positive target, the profit curve improves over the first percentiles, which means that the model will detect the population with the highest scores.
If you expand the node with the lowest percentage of positive target, the profit curve will improve over the last percentiles.

However, if the node you expand contains a very small population, the profit curve will not be impacted. So you need to find the best compromise between the size of the population and the percentage of positive target.
Customizing the Display

The button Display Settings allows you to customize some of the display settings for the decision tree.

- **Orientation**: this setting allows you to select if you want to display the tree horizontally or vertically.
- **Display Type**: this setting allows you to display the decision tree as a standard decision tree (Compact) or with a specific look provided by SAP BusinessObjects Predictive Analytics (Full). The Compact option uses less screen space, but the Full option is easier to read.

When you have set the display parameters, click the Close button.

5.5.4 Step 4 - Using the Model

Once generated, a classification model may be saved for later use.

A classification model may be applied to additional datasets. The model thus allows you to perform predictions on these application datasets, by predicting the values of a target variable. The model can also be used to carry out simulations on specific observations, on a case-by-case basis.

Moreover, you can refine a classification model, by re-generating it with an optimized list of explanatory variables. The application allows you to select the variables most pertinent to your business issue automatically, with pertinence defined as: producing the minimum area between the predictive curve and the hypothetical perfect curve, and thus maximizing the volume of information explained by the model.

So that you apply the model to any other database, the application allows you to generate different source code of the model (for example, C, XML, AWK, HTML, SQL, PMML2, SAS, and JAVA).

5.5.4.1 Analyzing Deviations

The option Analyze Deviations is a tool that provides you with a diagnostic of the data statistical variation.

This option can be used for several purposes:

- to compare the distribution of a new dataset with the distribution of the dataset used to train the model,
- to check the quality of new data after loading them,
- to check if your data have evolved over time and thus if the model need to be adapted to the new data.

5.5.4.1.1 Selecting the Dataset to Analyze

First, you need to select the dataset for which you want to analyze the deviations.

For the results to make sense, the new dataset should contain the same columns as the dataset that was originally used to train the model including the target variable, which must be filled.
To Select a Dataset

1. On the screen *Analyze Deviations*, select the data source format to be used ( *Text file*, *ODBC*, ...).
2. Click the *Browse* button.
3. In the *Data Selection* dialog, browse to the <Installation Path>/Samples folder.
4. Select the file you want to use, then click *OK*. The name of the file appears in the *Data Set* field.
5. Click *Next*. The screen *Deviation Analysis Debriefing* is displayed.

5.5.4.1.2 Following the Deviation Analysis Progress

The panel *Deviation Analysis Debriefing* allows you to follow the analysis process thanks to a progression bar. At the end of the process, a debriefing panel is displayed. For details on the debriefing panel, see the topic on Understanding the Deviation Analysis.

You can use the toolbar provided on the upper part of the panel to:

- stop the analysis process, by clicking the button,
- display the text log detailing the process, by clicking the button,
- copy, print or save the debriefing panel.

To Copy the Report

Click the (Copy) button.

The application copies the HTML code of the screen that you can the paste into another document (for example, a spreadsheet or text editor).

To Print the Report

1. Click the (Print) button situated under the title. A dialog box appears, allowing you to select the printer to use.
2. Select the printer to use and set other print properties if need be.
3. Click *OK*. The report is printed.

To Save the Report

Click the (Save) button situated under the title.
The file is saved in HTML format.

Related Information

Understanding the Deviation Analysis [page 133]

5.5.4.1.3 Understanding the Deviation Analysis

The first step to take to know if there are any deviations in your data is to look at the debriefing report and compare the performances (KI and KR) obtained on the original data with those obtained on the control dataset.

Then to visualize which variables have changed, you should look into the Control for Deviation Reports.

Debriefing Report

The section Control for Deviation Overview provides you with basic statistics on the Data Set used for Deviation Control (also called control dataset) such as:

- the name of the dataset (Data Set),
- the source file (Source),
- the number of records contained in the dataset (Number of Records)
- and the number of variables for which the application has found deviations in comparison to the dataset originally used to train the model (Number of variables showing deviation).

The second and third section of the debriefing report allows you to compare the performance of your model on the original dataset with the its performance on the control dataset:

- the section Performance Indicators displays for each target the KI and KR indicators obtained by the model on the original dataset.
- the section Performance on Control Data Set displays for each target the KI and KR indicators obtained by the model on the control dataset.

If the KI and/or KR of the model on the control dataset are significantly lower, it means that the relation between the variables and the target variable has changed. As a consequence the model should be rebuilt on the new data.

If the KI and KR are not much different, it means that the relation between the input variables and the target behavior has not changed, but it does not mean that differences of distributions are not possible.

Control for Deviations Reports

The panel Control for Deviations provides you with six options that can be separated in three groups:
• The first one, made of the options *Probability of Deviation*, *Probability of Category Deviation* and *Probability of Grouped Category Deviation*, enumerates the probabilities of deviation of each variable distribution, be it by variable, variable category or group of categories. A probability over 0.95 indicates that the variable or category global distribution is significantly different in the control dataset than in the reference dataset.

**Note**
The probability of deviation is actually a standardized chi-square test. It is significant above 0.95.

• The second group, comprised of the options *Probability of Target Deviation* and *Probability of Target Deviation for Grouped Categories*, lists for each variable the probabilities of deviation of the categories and the grouped categories with respect to the target variable. A probability over 0.95 indicates that there is a change of behavior with respect to the target variable in the category or group of categories.

• The last group contains only the option *Category with Problem*. For each dataset (reference datasets and control dataset), all variable categories with a probability over 0.95 are listed. This allows you a quick visualization of possible problems without having to analyze all the reports.

**Caution**
In all the report panels the control dataset is referred as the *ApplyIn* dataset.

**Options**

You can select which report sections to save:

1. Click the button **Save the reports** (located in the bottom left corner). A selector window opens.
2. In the displayed list, check the sections you want to save.
3. In the list **Report Style**, select the type of output you want. Three styles of output are available:
   - **Automatic**: saves the default view displayed in the interface
   - **Graphical**: saves the reports as graphs if such a view exists
   - **Textual**: saves the reports as tables

**Caution**
When selecting the options Automatic or Graphical, be careful to choose an appropriate file type such as pdf, rtf or HTML.

4. Click **OK**.
5. Select the folder in which you want to save the report.
6. Enter the name of the file.

### 5.5.4.2 Applying the Model to a New Dataset

The currently open model may be applied to additional datasets. The model allows you to perform predictions using the application datasets, and specifically, to predict the values of the target variable.
Constraints of Model Use

In order to apply a model to a dataset, the format of the application dataset must be identical to that of the training dataset used to generate the model. The same target variable, in particular, must be included in both datasets, even if the values are not contained in the application dataset.

**Note**

If the \textit{KxIndex} variable of the model is virtual, the application dataset must not contain a physical \textit{KxIndex} variable.

### 5.5.4.2.1 To Apply the Model to a New Dataset

For this scenario, due to technical constraints, a dataset corresponding to the database of 1,000,000 customers that will be used in this scenario can not be provided to you. You will apply the model to the file \texttt{Census01.csv}, which you used to generate the model. In this manner, you will be able to compare the predictions provided by the model to the real values of the target variable \textit{Class} for each of the observations.

In the procedure:

- Select the format \textit{Text files},
- In the \textit{Generate} field, select the option \textit{Individual Contributions},
- Select the folder of your choice in which to save the results file (\textit{Model-Generated Output}).
- Do not select the option \textit{Keep only outliers}.

1. On the screen \textit{Using the Model}, click the option \textit{Applying the model to a new dataset}.

   The screen \textit{Applying the Model} appears.

2. In the section \textit{Application Data Set}, select the format of the data source in the list \textit{Data Type}.

3. Click the \textit{Browse} button to select:
   - In the \textit{Folder} field, the folder which contains your dataset.
   - In the \textit{Data} field, the name of the file corresponding to your dataset.

4. In the section \textit{Results generated by the model}, select the file format for the output file in the list \textit{Data Type}.

5. You may also opt to select \textit{Keep only outliers}.

   If you select this option, only the outlier observations will be presented in the results file obtained, after applying a model.

6. Click the \textit{Apply} button.

   The screen \textit{Applying the Model} appears.

Once application of the model has been completed, the results files of the application is automatically saved in the location that you had defined from the screen \textit{Applying the Model}.
5.5.4.2.2 Classification Decision

The screen Classification Decision allows you to select how many observations you want the model to detect after application on the new dataset.

To apply a Classification Decision

1. On the screen Applying the Model, follow all the steps of the procedure To Apply a Model to a New Dataset.
2. In the Generate drop-down list, select the option Decision.
3. Click the Apply button.
   The screen Classification Decision appears.
4. Use the slide to set the percentage of population to detect.
5. Click Next.
   The model is applied to the new dataset.

Understanding the Classification Decision Screen

The screen Classification Decision allows you to either select a percentage of the population who will respond positively to your campaign (% of Detected Target) or a percentage of the entire population (% of Population).

When moving the cursor on the scale, the different values are updated accordingly.

For example, if you select the option % of Detected Target and set the cursor to 80%, the value of the field % of Population will be 32.0, which means that if you want that 80% of the people who will respond positively to your campaign receive your mailing, you will have to send it to 32% of the entire population.

On the other hand, if you select the option % of Population and set the cursor to 20% on the scale, the value of the field % of Detected Target will be 60.4, which means that if your budget only allows you to send your mailing to 20% of the entire population, you will touch 60% of the population who will respond positively.

For more details on how to use the Confusion Matrix, see the related topic.

Related Information

Confusion Matrix [page 123]

5.5.4.2.3 Using the Option Direct Apply in the Database

This optimized scoring mode can be used if all the following conditions are met:

- the apply-in dataset (table, view, select statement, data manipulation) and the results dataset are tables coming from the same database.
● the model has been computed while at least one physical key variable was defined in the application,
● no error has occurred,
● the in-database apply mode is not deactivated,
● granted access to read and write (create table).

To use the in-database apply mode, check the option *Use the Direct Apply in the Database* and automatically the option *Add Score Deviation* is selected as well.

5.5.4.2.4  Advanced Apply Settings

5.5.4.2.4.1  General Outputs

**Copy the Weight Variable**

This option allows you to add to the output file the weight variable if it had been set during the variable selection of the model.

**Copy Data Set Id**

This option allows you to add to the output file the name of the sub-data set the record comes from (*Estimation, Validation* or *Test*).

**Note**

This option cannot be used with the in-database apply feature.

**Copy the Variables**

This option allows you to add to the output file one or more variables from the data set.

- To Add All the Variables
  - Check the *All* option.
- To Select only Specific Variables
  1. Check the *Individual* option.
  2. Click the >> button to display the variable selection table.
  3. In the *Available* list select the variables you want to add (use the *Ctrl* key to select more than one variable).
  4. Click the > button to add the selected variables to the *Selected* list.
User Defined Constant Outputs

This option allows you to add to the output file constants such as the apply date, the data set name, or any other information useful for using the output file.

A user defined constant is made of the following information:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Value / Warnings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visibility</td>
<td>indicates if the constant will appear in the output or not</td>
<td><strong>checked:</strong> the constant appears in the output&lt;br&gt;<strong>unchecked:</strong> the constant does not appear in the output</td>
</tr>
<tr>
<td>Name</td>
<td>the name of the user defined constant</td>
<td>● The name cannot be the same as the name of an existing variable of the reference data set.&lt;br&gt;● If the name is the same as an already existing user defined constant, the new constant will replace the previous one</td>
</tr>
<tr>
<td>Storage</td>
<td>the constant type (number, string, integer, date, ...)</td>
<td><strong>number</strong>&lt;br&gt;<strong>string</strong>&lt;br&gt;<strong>integer</strong>&lt;br&gt;<strong>date</strong>&lt;br&gt;<strong>datetime</strong></td>
</tr>
<tr>
<td>Value</td>
<td>the value of the constant</td>
<td>date format: YYYY-MM-DD&lt;br&gt;datetime format: YYYY-MM-DD HH:MM:SS</td>
</tr>
<tr>
<td>Key</td>
<td>indicates if the constant is a key variable or identifier for the record. You can declare multiple keys. They will be built according to the indicated order (1-2-3-...).</td>
<td>0: the variable is not an identifier&lt;br&gt;1: primary identifier&lt;br&gt;2: secondary identifier&lt;br&gt;...</td>
</tr>
</tbody>
</table>

- To Define a Constant
  1. Click the **Add** button. A pop-up window opens allowing you to set the constant parameters.
  2. In the field **Output Name**, enter the constant name.
  3. In the list **Output Storage**, select the constant type.
  4. In the field **Output Value**, enter the constant value.
  5. Click the **OK** button to create the constant. The new constant appears in the list. You can choose whether to generate the defined constants or not by checking the **Visibility** box.
5.5.4.2.4.2 Gain Chart

This tab allows you to compute the gain chart on the apply dataset, that is to rank your data in order of descending scores and split it into exact quantiles (decile, vingtile, percentile).

If you have computed the gain chart while creating your model, two gain charts will be computed during the application process:
- a transversal gain chart allowing you to check the frequency deviation between the validation and the application gain chart,
- an apply gain chart allowing you to have the exact number of targets inside each tiles.

1. Check the box **Compute Gain Chart on Apply-in Data**.
2. In the list, select the **Number of Quantiles** you want your data to be segmented in.
3. You can add additional variables in order to estimate profits per segments of the population:
   a. In the **Variables** list, select the variables you want to add to the gain chart. Use the CTRL key to select multiple variables.
   b. Click the > button to add the selected variables to the list **Values for Gain Chart**.
4. The sum of each selected variable will be calculated for each segment of the population.
5. Click **Validate** to save the advanced parameters and go back to the panel **Applying the Model**.

The result of the gain chart computation is available at the end of the model application. It can also be found in the **Statistical Reports**, in the section **Model Performance**.

If several gain charts have been computed, select the dataset in the proposed list to display the gain chart you want to visualize.

5.5.4.2.4.3 Outputs by Targets

5.5.4.2.4.3.1 Reason Codes

Reason codes are variables whose values have the most influence in a score-based decision (typically a risk score). An example of the use of reason codes is to provide a customer with the reasons why the automatic scoring system did not approve his loan.

To Generate Reason Codes

1. In the tree **Advanced Apply Settings** located on the left of the panel, open the node **Outputs for Target** <Target Name>.
2. Select **Reason Codes**.
3. Click the + button located on the right of the displayed table.
4. Click in the cell corresponding to the parameter you want to set. The following table sums up the available parameters:
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Values</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Reason Codes</td>
<td>Integer</td>
<td>Number of reason codes you want to generate.</td>
</tr>
<tr>
<td></td>
<td>Default: 3</td>
<td></td>
</tr>
<tr>
<td>Threshold</td>
<td></td>
<td>Threshold used for computing the most important reason codes. For each variable the contribution corresponding to the customer score is compared to its contribution for the whole population. The variables for which the contribution is the most differential are selected as the most important reason codes. For example, if you select Mean, the customer variable contribution will be compared to the mean of the whole population contribution to determine which variables are the most differential.</td>
</tr>
<tr>
<td></td>
<td>○ Mean (default)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>○ Maximum</td>
<td></td>
</tr>
<tr>
<td></td>
<td>○ Minimum</td>
<td></td>
</tr>
<tr>
<td>Criterion</td>
<td>○ Below (default)</td>
<td>Indicates whether you want to generate the reason codes when the customer variable contribution is above or below the threshold. Warning: Using Below with the Minimum threshold or Above with the Maximum threshold will generate an error.</td>
</tr>
<tr>
<td></td>
<td>○ Above</td>
<td></td>
</tr>
</tbody>
</table>

5. If you want to generate several types of reason codes, repeat steps 3 and 4 for each type.

Output

The output table contains two columns for each reason code requested:

- `reason_name_<criterion>_<threshold>_<rank>_rr_<target name>`: contains the name of the variable selected as a reason code. For example, the output column named `reason_name_Below_Mean_1_rr_class` contains the name of the variable being the most important (1) reason code with respect to the target variable class. Among the variables whose contribution is below (Below) the mean (Mean) of the population contribution, the selected variable will be the one having the highest deviation with it.

- `reason_value_<criterion>_<threshold>_<rank>_rr_<target name>`: contains the value of the reason code.
## 5.5.4.2.4.3.2 Continuous Target

<table>
<thead>
<tr>
<th>Option</th>
<th>Output Column Name</th>
<th>This option allows you to...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted Value</td>
<td>rr_&lt;target variable&gt;</td>
<td>generate in the output file the value predicted by the model for the target variable. This option is checked by default.</td>
</tr>
</tbody>
</table>
| Confidence          | bar_rr_<target variable> | add to the output file the confidence level for the value that has been predicted, this is also known as the error bar. It is computed with “3” standard deviations on the validation dataset and bin per bin. The percentage of population corresponding to the “3” standard deviations is of about 99%. Calculation formula: \( \{ \text{TargetMean} - 3 \times \sqrt{\text{TargetVariance}}; \text{TargetMean} + 3 \times \sqrt{\text{TargetVariance}} \} \)
|                     |                          | When \( \sqrt{\text{TargetVariance}} \) is equal to the Standard Deviation \( \text{TargetMean} \pm \text{Standard Deviation} \) is equal to the Confidence Interval                                                                  |
| Outlier Indicator   | outlier_rr_<target variable> | to show in the output file which observations are outliers. An observation is considered an outlier if the difference between its "predicted value" and its "real value" exceeds the value of the error bar. In other words, the error bar is a deviation measure of the values around the predicted score. Possible values are 1 if the observation is an outlier with respect to the current target, else 0. |
| Contributions       | contrib_<variable>_rr_<target variable> | add the variables contributions for the current variable to the output file. You can add the contributions of all variables or select only the contributions of specific variables (see procedure below). |

Classification, Regression, Segmentation and Clustering Scenarios
Modeler - Regression/Classification
Predicted Value

This option is checked by default. It allows you to generate in the output file the value predicted by the model for the target variable. It appears in the output file as \texttt{rr_<target variable>}. 

Outlier Indicator

This option allows you to show in the output file which observations are outliers. An observation is considered an outlier if the difference between its "predicted value" and its "real value" exceeds the value of the error bar. In other words, the error bar is a deviation measure of the values around the predicted score. It appears in the output file as \texttt{outlier_rr_<target variable>}. Possible values are 1 if the observation is an outlier with respect to the current target, else 0.

Predicted Value Quantile

This option allows you to cut the output file in quantiles and to assign to each observation the number of the quantile containing it.

Approximate quantiles are constructed based on the sorted distribution and the boundaries of predicted scores from the validation sample. The score boundaries are used to determine approximate quantiles on the apply data set.

\begin{itemize}
  \item Exact quantile computation would require a full sort of the scores obtained on the apply data set which can be consuming.
  \item A Gain Chart option is available for this purpose.
\end{itemize}

It appears in the output file as \texttt{quantile_rr_<target variable>_<number of quantiles>}, for example for a target variable named "class" and a number of quantiles equal to 10, the generated column will be named \texttt{quantile_rr_class_10}.

1. Check the option \textit{Predicted Value Quantiles}.
2. In the field \textit{Number of Quantiles}, enter the number of quantiles you want to create. Check the option \textit{Predicted Value Quantiles}.

Contributions

This option allows you to add the variables contributions for the current variable to the output file. You can add the contributions of all variables or select only the contributions of specific variables.

It appears in the output file as \texttt{contrib_<variable>_rr_<target variable>}. 

For example, if "marital-status" is an explanatory variable for the target variable "class", the column `contrib_marital-status_rr_class` will be generated in the output file.

- To Add All Variables Contributions
  - Check the All option.
- To Add Specific Variable Contributions
  1. Check the Individual option.
  2. Click the >> button to display the variable selection table.
  3. In the Available list select the variables you want to add (use the Ctrl key to select more than one variable).
  4. Click the > button to add the selected variables to the Selected list.

5.5.4.2.4.3.3 Nominal Target

Outputs by Rank

Scores

This option allows you to generate in the output file the best score(s) for each observation. For each line in the application data set, Automated Analytics compares the scores obtained by the current observation for each category of the target variable and displays the best score in the column `best_rr_Target Variable>_1`. You can generate up to 2 best scores since a target variable has 2 categories. The second best score is displayed in the column `best_rr_Target Variable>_2`. When using this option with the Decision option described below, you can link the best score with the category that has obtained it.

Decision

This option allows you to generate in the output file the best decision(s) for each observation. Like for the previous option, the scores obtained for each category of the target variable are compared and the category with the best score for the current record is displayed in the column `decision_rr_Target Variable>`. You can generate up to 2 best decisions since a target variable has 2 categories. The category with the second best score is displayed in the column `decision_rr_Target Variable>_2`.

Probabilities

This option allows you to generate in the output file the probability of the best decisions for each observation. Like for the previous options, the scores obtained for each category of the target variable are compared and the probability of the category with the best score for the current record is displayed in the column `proba_rr_Target Variable>`. You can generate up to 2 probabilities of the best decisions since a target variable has 2 categories. The probability of the category with the second best score is displayed in the column `proba_rr_Target Variable>_2`.
Outputs by Reference Category

Score

This option allows you to generate in the output file the score corresponding to each data set line for the different categories of the target variable. You can generate the scores for all the target variable categories or select specific categories.

It appears in the output file as \texttt{rr_<Target Variable>\_<Category>} for the target variable key category and \texttt{rr_<Target Variable>\_<Category>} for its other categories.

- To Add the Score of All Target Variable Categories
  - Check the \textit{All} option.
- To Add Only the Scores of Selected Categories
  1. Check the \textit{Individual} option.
  2. In the \textit{Selection} column, check the boxes corresponding to the categories for which you want to add the score in the output file.

Prediction Probability

This option allows you to generate in the output file the probability for one or more target variable categories, that is for each observation the probability of the target variable value to be the selected category.

It appears in the output file as \texttt{proba_rr_<Target Variable>\_<Category>} for the target variable key category and as \texttt{proba_rr_<Target Variable>\_<Category>} for the other categories of the target variable.

- To Add the Probabilities of All Target Variable Categories
  - Check the \textit{All} option.
- To Add Only the Probabilities of Selected Categories
  1. Check the \textit{Individual} option.
  2. In the \textit{Selection} column, check the boxes corresponding to the categories for which you want to add the probabilities in the output file.

Miscellaneous Outputs

Outlier Indicator

This option allows you to show in the output file which observations are outliers. An observation is considered an outlier if the difference between its "predicted value" and its "real value" exceeds the value of the error bar. In other words, the error bar is a deviation measure of the values around the predicted score. It appears in the output file as \texttt{outlier rr\_<target variable>}. Possible values are 1 if the observation is an outlier with respect to the current target, else 0.
Predicted Value Quantile

This option allows you to cut the output file in quantiles and to assign to each observation the number of the quantile containing it.

*Approximate* quantiles are constructed based on the sorted distribution and the boundaries of predicted scores from the validation sample. The score boundaries are used to determine *approximate* quantiles on the apply data set.

**Note**

- Exact quantile computation would require a full sort of the scores obtained on the apply data set which can be consuming.
- A *Gain Chart* option is available for this purpose.

It appears in the output file as `quantile_rr_<target variable>_<number of quantiles>`, for example for a target variable named "class" and a number of quantiles equal to 10, the generated column will be named `quantile_rr_class_10`.

1. Check the option *Predicted Value Quantiles*.
2. In the field *Number of Quantiles*, enter the number of quantiles you want to create.

Contributions

This option allows you to add the variables contributions for the current variable to the output file. You can add the contributions of all variables or select only the contributions of specific variables.

It appears in the output file as `contrib_<variable>_rr_<target variable>`.

For example, if "marital-status" is an explanatory variable for the target variable "class", the column `contrib_marital-status_rr_class` will be generated in the output file.

- To Add All Variables Contributions
  - Check the *All* option.
- To Add Specific Variable Contributions
  1. Check the *Individual* option.
  2. Click the >> button to display the variable selection table.
  3. In the *Available list* select the variables you want to add (use the [Ctrl] key to select more than one variable).
  4. Click the > button to add the selected variables to the *Selected list*.

5.5.4.2.5 Types of Results Available

The application of a model to a dataset allows you to obtain four types of results, which are described in the following table.
### Type of Results

<table>
<thead>
<tr>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>score, or predicted value</td>
</tr>
<tr>
<td>probability</td>
</tr>
<tr>
<td>prediction range, or maximum error</td>
</tr>
<tr>
<td>individual contributions</td>
</tr>
<tr>
<td>decision</td>
</tr>
</tbody>
</table>

Upon the level of information desired, you can choose to generate among several results' files, described in the table below.

<table>
<thead>
<tr>
<th>Selecting the option...</th>
<th>Will generate a results' file containing the following information...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted value only</td>
<td>Only the predicted value of observations (rr_&lt;target variable name&gt;)</td>
</tr>
</tbody>
</table>
| Probability             | • the predicted value  
                          | • the probability (proba_rr_<target variable name>)  
                          | • the prediction range (bar_rr_<target variable name>) |
Selecting the option... | Will generate a results' file containing the following information...

| Individual Contributions | • the predicted value  
|                        | • the probability  
|                        | • the prediction range  
|                        | • the individual contributions of variables  
|                        | (contrib_VariableName_rr_<target variable name>)  

| Decision | • the predicted value  
|          | • the decision(decision_rr_<target variable name>)  
|          | • the decision probability(proba_decision_rr_<target variable name>)  
|          | • the probability  

**Related Information**

To Apply the Model to a New Dataset [page 135]

### 5.5.4.2.5.1 Analyzing the Results of the Application

For this scenario, open the results file in Microsoft Excel, in the text format that you obtained when you applied the model to the Census01.csv file.

To Open the Model Application Results File

1. Depending upon the format of the results file generated, use Microsoft Excel or another application to open the file. The figure below presents the headings and columns of the results file obtained for this scenario.

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Kfold</td>
<td>class</td>
<td>proba_class</td>
<td>bar_class</td>
<td>contrib_age</td>
<td>contrib_workclass</td>
</tr>
</tbody>
</table>
| 2  | 1     | 0.0262205544  | 0.17731641 | 1.17675617 | 0.0001418 | 0.004579982  
| 3  | 2     | 0.3687523033  | 0.007958506 | 1.456298657 | 0.00034662 | 0.004579982  
| 4  | 3     | 0.1963907756  | 0.022777327 | 0.457114785 | 0.00033685 | -0.002216466  
| 5  | 4     | 0.032755154     | 0.35421767 | 0.465407827 | 0.00075412 | -0.002216466  
| 6  | 5     | 0.310153429     | 0.549035046 | 1.261932184 | 0.000754755 | -0.002216466  
| 7  | 6     | 0.35425714      | 0.773501855 | 1.401859995 | 0.000323156 | -0.002216466  
| 8  | 7     | 0.360559064     | 0.00035657 | 0.461765275 | 0.000503965 | -0.002216466  
| 9  | 8     | 0.1800327327    | 0.439521527 | 1.493381023 | 0.000711806 | 0.004579982  
| 10 | 9     | 0.034126425     | 0.022523628 | 0.592684474 | 0.00133935 | -0.002216466  
| 11 | 10    | 0.540310015     | 0.162732898 | 0.903858706 | 0.001757386 | -0.002216466  
| 12 | 11    | 0.356238005     | 0.515526811 | 1.519433461 | 0.000221956 | -0.002216466  
| 13 | 12    | 0.3177735808    | 0.074707566 | 1.435375769 | 0.000437512 | 0.004579982  
| 14 | 13    | 0.152527474     | 0.024235664 | 0.406258554 | 0.001327858 | -0.002216466  
| 15 | 14    | 0.11625945      | 0.029107656 | 0.751248468 | 0.00031325 | -0.002216466  
| 16 | 15    | 0.164723878     | 0.40374312 | 1.466822715 | 0.00038525 | -0.002216466  

2. You can now analyze the results obtained and use these results of your analysis to make the right decisions.
Description of the Results File

Depending upon which options you selected, the results file will contain some or all of the following information, in the same order as seen below:

- The key variable defined during data description at the setting model parameters step.
- Possibly, the target variable given as known values if the latter appeared in the application dataset, as is the case in this scenario.
- The predicted value (score) provided by the model for the target variable of each observation. The name of this column corresponds to the name of the target variable prefixed by rr_, or in this case, rr_Class.
- The decision is based on the score. For example, its value can be of 1 if the observation is considered as interesting or 0 if it is considered as uninteresting for the model. The name of this column corresponds to the name of the target variable prefixed by decision_rr_, or in this case, decision_rr_class.
- The probability decision is also based on the score and provides the probability of the decision. The higher it is, the more it will confirm the decision value. The name of this column corresponds to the name of the target variable prefixed by proba_decision_rr_, or in this case, proba_decision_rr_class.
- The probability for each observation that it does or does not belong to the target category of the target variable. The name of this column corresponds to the name of the target variable prefixed by proba_rr_, or in this case, proba_rr_class.
- The prediction range, or "maximum error". The name of this column corresponds to the name of the target variable prefixed by bar_rr_, or in this case, bar_rr_Class.
- The individual contributions by variables contained in the dataset with respect to the target variable. The names of the columns of individual contributions correspond to the names of each of the variables, prefixed by contrib_, or in this case, contrib_age, contrib_workclass, and so on.

5.5.4.3 Performing a Simulation

The open model may be used to carry out simulations on specific observations, one at a time. To define the observation to be analyzed, the variables of your choice must be associated with values. For instance, if you have selected the occupation (profession category) and workclass (socio-professional category) variables, they must contain values. During execution of the simulation, the application will automatically assign values to certain variables when values are missing, but essential to proper completion of the simulation.

Once the simulation is complete, you will obtain the following results:

- The predicted value (score).
- The probability that this observation belongs to the target category of the target variable.

To Simulate a Model

1. On the screen Using the Model, click the option Simulation.

   The screen Simulating the Model appears.

2. On the left side of the screen (Explanatory variables), select a variable, such as marital-status. Its values appear in the section Modifying values, on the right side of the screen.
3. In the section *Modifying values*, in the *Value* field, select or enter a value, such as *Married-civ-spouse*. The value appears in the table of *Explanatory variables*, across from the selected variable.
4. If you would like to select other explanatory variables, go back to step 2. Otherwise, go to step 5.

5. Click the Run button to perform a model simulation.

The results of the simulation appear in the Results section. You will obtain the predicted value (score) of the observation described in the table of Explanatory variables, as well as the probability that this observation belongs to the target category of the target variable. In our example, only one variable has been defined. The probability that this observation belongs to the target category of the target variable is 0.1120.

Note that certain variables of the table of Explanatory variables were automatically completed upon execution of the simulation. In fact, the model automatically completed certain missing values that were essential to the simulation.

These values are listed in the following table.

<table>
<thead>
<tr>
<th>Type of variable</th>
<th>Default value</th>
</tr>
</thead>
<tbody>
<tr>
<td>continuous variable</td>
<td>the mean value</td>
</tr>
<tr>
<td>nominal variable</td>
<td>the most frequent category</td>
</tr>
<tr>
<td>ordinal variable</td>
<td>the most frequent category</td>
</tr>
</tbody>
</table>

These changes are reflected in the left part of the screen after clicking the Run button.
6. You can modify the value of an explanatory variable and run the simulation again to measure the effect of that change with respect to the target variable. For instance:
   a. Assign the value Widowed to the variable marital-status in place of the value Married-civ-spouse.
   b. Run the simulation. The probability now obtained is 0.0040.

7. Click the Reset button to run the simulation again.

### 5.5.4.4 Refining a Model

The application allows you to refine a currently open model. For instance, you can:

- Reduce the number of explanatory variables used by the model while maintaining the initial quality (KI) and robustness (KR).
- Generate a model of degree 2 using the most significant variables of the degree 1 model.

#### Note

If your dataset contains date or datetime variables, automatically generated variables will appear in this panel. For more information, refer to the information on date and datetime variables in the topic on variable storage formats.

1. On the screen Using the Model, click the option Select Variables.

   The screen Selecting Contributory Variables appears.
2. In the **Targets** list, select the target variable for which you want to select the contributory variables.

3. Click the button **Smart Selection**. The window **Smart Variables Selection** opens.

4. On the bar **Percentage of Information Retained**, move the cursor to change the amount of information to keep; the number of variables selected changes accordingly.

   The further this cursor is moved to the left, the more variables are excluded. The variables excluded are selected automatically as a function of their significance with respect to the model. For instance, the figure below shows that to retain only two variables out of the original fourteen, you should keep 43.07% of the information contributed by the model.

```
Note

Certain variables in the training dataset may contribute no information, such as constant value variables. These can therefore be automatically excluded from the model during the training phase. The number of variables excluded is displayed as a **Remark**. In the figure above, this number is equal to "0".
```

5. Click **OK**.

   The window **Smart Variables Selection** closes and the panel **Selecting Contributory Variables** is updated with the selected variables, allowing you to view the kept variables and the excluded ones. In our example, the application automatically determined that the two explanatory variables that contributed the most information to explain the target variable were the variables **marital-status** and **capital-gain**.
6. Click Next.

A message "This will reset the current. Do you really want to do this?" appears.

7. Click Yes to move to the screen Selecting Variables.

8. Resume the model configuration from the step selecting variables.

Related Information

Variable Storage Formats [page 30]
Selecting Variables [page 80]

5.5.4.5 Generating the Source Code of a Model

1. In the list Target to be used, choose the target of model.
2. Use the list Information to be generated to select the type of results:
Selected Option | Results of the Generated Model
--- | ---
Score/Estimates | score value (classification) or estimates (regression)
Probability | score value and probability value, except for HTML and all SQL codes for which only the probability value is provided.
Bar | score value and error bar value, except for HTML and all SQL codes for which only the error bar value is provided.

**Caution**

Both options Probability and Bar are only available for regression/classification models with nominal targets.

**Note**

In the case of a continuous variable, the generated code (SQL for example) always includes a number of categories that is higher than in the user defined structure or as given by the parameter band count if no user has structure has been set. Indeed, the encoding of variables adds target curve points to increase the accuracy of coding according to the training dataset. These points split some existing categories and thus increase the number of categories in the generated code.

3. In the section **Code Settings**, select the code type to be generated.
4. Click the **Browse** button associated with the **Folder** field and select a folder to save the generated file.
5. In the field **Generated File**, enter the name of the exported file. If you want to replace an existing file, use the **Browse** button to select it.
6. If you have selected the option **View Generated Code**, it is displayed at the end of the generation process.
7. Click the **Generate** button.

**Related Information**

List of Generated Codes [page 154]

### 5.5.4.5.1 List of Generated Codes

The following table lists the available codes with their particularities:

<table>
<thead>
<tr>
<th>Generated Code</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>see C Code Generator documentation</td>
</tr>
<tr>
<td>JAVA</td>
<td>needs the KxJRT.jar package to run</td>
</tr>
<tr>
<td>PMML3.2</td>
<td></td>
</tr>
<tr>
<td>AWK</td>
<td></td>
</tr>
</tbody>
</table>
**Generated Code** | **Comment**
--- | ---
CPP |  
SAS |  
SQLServer | wraps variable names with [ ]
SQLServerUDF |  
HANA | the code generator manages SAP HANA column and row storage
ORACLE |  
OracleUDF |  
SQLDB2 |  
DB2UDF |  
DB2V9 |  
SQLTeradata | Teradata databases
TERAUDF |  
MYSQL |  
MYSQLUDF |  
SybaseIQ |  
SybaseIQUDF |  
SQLNetezza |  
SQLVertica |  
PostgreSQL |  
Greenplum |  
Hive |  
HTML (Javascript) | contains a form to fill which reproduces the model
ScoreCard | only available for regression/classification models
CCL Code | for scoring in SAP HANA Smart Data Streaming

**Note**
When generating SQL and SAS codes, you will be asked to provide the names of the key column and of the dataset used.

**Caution**
Only SQLServer key code handles trimmed data during its execution. For other codes, if data are not trimmed it may generate some differences.
5.5.4.5.2 Advanced Settings

UNICODE Mode

The option Activate UNICODE Mode allows you to generate the code selected in Unicode so that it supports non-latin languages such as Japanese, Russian, and so on. This option is particularly useful for SQL codes.

SQL/UDF Options

- The option Do not generate code for non-contributive variables allows you to exclude from the code all variables with a contribution of 0 since they do not influence the result. In some cases, this can significantly reduce the size of the generated code.
- You can either Use the default separator ("GO") or Use a custom separator.

5.5.4.5.3 Exporting the Model as a KxShell Script

The KxShell script export allows you to generate a KxShell script reproducing the current model. This script can be used to run models in batches.

One easy way to get special settings in exported KxShell scripts is to first do the corresponding operation in the graphical user interface. For example, if you run an auto-selection of variables before exporting the shell script, then the exported script will include the code needed to do the auto-reduction.

To save the KxShell script:

1. In the section Save/Export of the menu Using the Model, select the option Export KxShell Script.

   The panel KxShell Script Generation is displayed.

2. Use the Browse button located to the right of the Folder field to select where the script will be saved.

3. In the field KxShell Script, enter the name of the file in which the script will be saved.

4. In the frame Model Data Set Description Saving, select where you want to save the data description. The four available options are:

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Save the Description in the Script</td>
<td>The data description is added in the KxShell script. Only one file is generated.</td>
</tr>
<tr>
<td>Save the Description with the Script</td>
<td>The data description is saved in an additional file in the same folder as the KxShell script.</td>
</tr>
<tr>
<td>Save the Description with the Data</td>
<td>The data description is saved in an additional file in the same folder as data used for the model.</td>
</tr>
</tbody>
</table>
### Option Description

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Save the Description Separately</td>
<td>The data description is saved in an additional file. Indicate the type of the description (text file, data base, flat memory, ...) and the location where the data description should be saved.</td>
</tr>
</tbody>
</table>

**Note**

When saving the description in an additional file, the file is named following this syntax: KxDesc_<Dataset Role>_<Dataset Name>. For example, for a training dataset named Census.csv, the description file name will be KxDesc_Training_Census.csv.

5. Additionally you can export the variable structure with relation to a target variable by checking the option *Generate Variable Structure From Statistics* and selecting the target variable in the list *Select a Target*.

   This option allows you to force the grouping of categories when training the model on new datasets.

6. Before exporting the script you can view the script by clicking the button *Script Preview*.

7. Validate to start the generation process. Once the script has been generated, the menu *Using the Model* is displayed.

### 5.5.4.6 Saving the Model

Once a model has been generated, you can save it. Saving it preserves all the information that pertains to that model, that is, the modeling parameters, its profit curves, and so on.

1. On the screen *Using the Model*, click the option *Save Model*.
   
   The screen *Saving the Model* appears.

2. Complete the following fields:

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Model Name</strong></td>
<td>This field allows you to associate a name with the model. This name will then appear in the list of models to be offered when you open an existing model.</td>
</tr>
<tr>
<td><strong>Description</strong></td>
<td>This field allows you to enter the information of your choosing, such as the name of the training dataset used, the polynomial degree or the predictive power and prediction confidence obtained for the model. This information could be useful to you later for identifying your model</td>
</tr>
</tbody>
</table>
   | **Data Type** | This list allows you to select the type of storage in which you want to save your model. The following options are available:  
   | | ○ *Text files*, to save the model in a text file.  
   | | ○ *Database*, to save the model in a database.  
   | | ○ *Flat Memory*, to save the model in the active memory.  
   | | ○ *SAS Files*, to save the model in a SAS compatible file for a specified version of SAS and a specified platform (SAS v6 or 7/8 for Windows or UNIX).  
   | | ○ *SAS Transport*, to save the model in a generic SAS compatible file:  
   | | ○ *Folder*: Depending upon which option you selected, this field allows you to specify the ODBC source, the memory store or the folder in which you want to save the model. |
Option | Description
--- | ---
○ **File/Table**: This field allows you to enter the name of the file or table that is to contain the model. When saving as a text file, you must enter one of the following format extensions. `.txt` (text file in which the data is separated by tabs) or `.csv` (text file in which the data is separated by commas).

### Related Information

**Files Created When Saving a Model**

When saving a model, the application creates a set of files/tables in the specified store. Some of these files are specific to the type of model. The following table lists the files or tables created when saving a model and in which case.

<table>
<thead>
<tr>
<th>File</th>
<th>Description</th>
<th>Used By</th>
</tr>
</thead>
<tbody>
<tr>
<td>KxAdmin</td>
<td>Lists all the models contained in the folder/database with additional information (date, version, name of the model, comments).</td>
<td>All models created with Automated Analytics</td>
</tr>
<tr>
<td><code>&lt;Model_name&gt;</code></td>
<td>File named after the model and containing all the model data, except graphs information. Graphs are stored in additional tables (see below).</td>
<td>All models created with Automated Analytics</td>
</tr>
<tr>
<td>KxInfos</td>
<td>Indicates which additional tables are needed by the model.</td>
<td>All models created with Automated Analytics</td>
</tr>
<tr>
<td>KxOlapCube</td>
<td>Stores the OLAP Cube used by the decision tree when the option Regression/Classification as Decision is activated.</td>
<td>Regression/Classification models with decision tree</td>
</tr>
<tr>
<td>KxLinks</td>
<td>Contains the links from the graphs of the model.</td>
<td>Social model only</td>
</tr>
<tr>
<td>KxNodes</td>
<td>Lists all the nodes from all the graphs and their attributes.</td>
<td>Social model only</td>
</tr>
<tr>
<td>KxCommunities</td>
<td>Matches the nodes to their communities, if the community detection was enabled.</td>
<td>Social model only</td>
</tr>
</tbody>
</table>

⚠️ **Caution**

When sharing or sending a model, all these files must be joined to the model or the recipient will not be able to open the model.
5.5.4.7 Opening a Model

Once saved, models may be opened and reused in the application.

1. On the main application screen, select **Load a Model**. The screen Opening a Model appears.
2. In the **Data Type** list, select one of the following options depending upon the format of the model that you want to open:
   - **Text files**
   - **Database**
   - **SAS files**
   - **SAS Transport**
3. Click the **Browse** button.
   
   The **Data Selection** dialog appears.
4. Select the folder that holds the model that you want to open. The list of models contained in that folder appears providing the following information for each model.

<table>
<thead>
<tr>
<th>Column</th>
<th>Description</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Name</strong></td>
<td>Name under which the model has been saved</td>
<td>Character string</td>
</tr>
<tr>
<td><strong>Class</strong></td>
<td>Class of the model, that is the type of the model</td>
<td>○ Kxen.Classification: Classification/Regression with nominal target</td>
</tr>
<tr>
<td></td>
<td></td>
<td>○ Kxen.Regression: Classification/Regression with continuous target</td>
</tr>
<tr>
<td></td>
<td></td>
<td>○ Kxen.Segmentation: Clustering with SQL Mode</td>
</tr>
<tr>
<td></td>
<td></td>
<td>○ Kxen.Clustering: Clustering without SQL Mode</td>
</tr>
<tr>
<td></td>
<td></td>
<td>○ Kxen.TimeSeries: Time Series</td>
</tr>
<tr>
<td></td>
<td></td>
<td>○ Kxen.AssociationRules: Association Rules</td>
</tr>
<tr>
<td></td>
<td></td>
<td>○ Kxen.SimpleModel: Classification/Regression and Clustering multi-target models, any other model</td>
</tr>
<tr>
<td><strong>Version</strong></td>
<td>Number of the model version when the model has been saved several times</td>
<td>Integer starting at 1</td>
</tr>
<tr>
<td><strong>Date</strong></td>
<td>Date when the model has been saved</td>
<td>Date and time in the format yyyy-mm-dd hh:mm:ss</td>
</tr>
<tr>
<td><strong>Comment</strong></td>
<td>Optional user defined comment that can be used to identify the model</td>
<td>Character string</td>
</tr>
</tbody>
</table>

5. Select a model from the list.
6. Click the **Open** button.

   The screen **Using the Model** appears.
6 Modeler - Segmentation/Clustering

6.1 Segmentation/Clustering Application Scenario Description

In this scenario, you are the Marketing Director of a large retail bank. You want to customize your communications using data modeling. The bank wants to offer a new financial product to its customers. Your project consists of launching a direct marketing campaign aimed at promoting this product.

In order to customize the marketing messages from the bank and improve communication with the various customers and prospects for this new product, the senior management of the bank asks you to build a segmentation model of the customers of this product.

Using Modeler - Segmentation/Clustering, you can rapidly develop a descriptive model with the least possible cost. This model shows the characteristic profiles of the customers interested in your new product, and thus responds to your business issue and fulfills your objectives.

Your Objective

Consider the following case.

Using Regression/Classification, you have contacted the prospects most likely to be interested in your new financial product, and identified the ideal number of prospects to contact out of the entire database meeting the deadlines and within the budget you were allowed.

To improve the rate of return of your campaign, senior management asks you to:

- Build a segmentation model of your customers,
- Analyze the characteristics of the identified clusters,
- Define customized communications for each cluster.

The segmentation model in particular should allow you to distinguish customer clusters by virtue of their propensity to purchase the new high-end savings product proposed by your firm. You will optimize your understanding of your customers.

Your Approach

For organizational reasons, you want to define five groups of customers, or clusters, and describe the customer profiles for each of these groups.

To accomplish this project, you will use the sample of 50,000 people who responded to your first test, during the previous campaign. This file corresponds to the sample file Census01.csv, provided with the application and described in the section Introduction to Sample Files.
Your Business Issue

In your marketing database, you have:

- A list of 1,000,000 prospects.
- A list of 50,000 prospects (people selected during the test phase of your campaign), whose responses to the campaign are known. This sample thus constitutes a training dataset. This sample, taken from the complete database, also exhibits some missing values.

Your business issue thus consists of:

- Rapidly building a segmentation model using the training dataset (or sample). The clusters obtained will allow you to better understand the profiles of the individuals in your database as a function of their propensity to purchase.
- Then applying the segmentation model obtained from the training data to the entire list of prospects to determine which cluster each individual should belong.

Related Information

Introduction to Sample Files [page 163]

6.2 Scenario Solutions

6.2.1 Intuitive Method

This method consists of using your knowledge of the various profiles exhibited by your customers. Thanks to the domain-specific knowledge that you have of your customers, you determine the criteria of the segmentation model intuitively, and build the clusters yourself.

The main disadvantage of this method is that the number of information items available for each customer will invariably grow with time. The more data your database accumulates, the harder it is for you to manually create clusters that take all data into consideration and to develop a response to your business issue. Furthermore, as the increasing volume of information requires you to build segmentation models with increasing frequency, the time required to build these segmentation models becomes increasingly more significant.

Finally, management may want you to rationalize your methods, and to perform your segmentation using a method not based purely on your intuition. Defending your segmentation method based on intuition may be difficult.
6.2.2 Classical Statistical Method

On the basis of the information that you have, a data mining expert could build a segmentation model. In other words, you could ask a statistical expert to create a mathematical model that would allow you to build clusters based on the profiles of your customers.

To implement this method, the statistician must:

- Perform a detailed analysis of your database.
- Prepare your database down to the smallest detail, specifically, encoding the variables as a function of their type (nominal, ordinal or continuous) in preparation for segmentation. The encoding strategy used will determine the type of segmentation model obtained. At this step, the statistician may unconsciously bias the results.
- Test different types of algorithms (K-means, both ascending and descending hierarchical segmentation models) and select the one best suited to your business issue.
- Evaluate the relevance of the clusters obtained, in particular, the response to your domain-specific business issue.

After a few weeks, the statistical expert will be able to provide a certain number of clusters, or homogeneous groups, to which each of the individuals of your database are assigned.

This method presents significant constraints. You must:

- Ensure that your statistical expert, who is usually from an external department, is available for the scheduled period,
- Ensure that the modeling costs will fit into your budget,
- Spend time explaining your domain-specific business issue to the statistician,
- Spend time understanding the results that are provided,
- Ask a programmer to write a program to determine the cluster associated with any new individual added to your database.

In addition, this method is not systematic. Two statisticians performing this segmentation on the same dataset could obtain different results.

6.2.3 Automated Analytics Method

Segmentation/Clustering allows you to build a segmentation model of your customers in a few minutes, taking into consideration the interest expressed by your customers in your new product.

Segmentation/Clustering automatically detects interactions between the variables to build homogeneous subsets, or clusters. Each cluster is homogeneous with respect to the entire set of variables, and in particular with respect to the target variable, that is, for example, "responded positively to my test".

You will discover the characteristics of different clusters, such as those clusters with an excellent response rate and those with a poor response rate. In addition, if your customer database contains customer expenditures on your other products, you will also obtain information on product sale synergies, by cluster.

Using Segmentation/Clustering, you have access to all the analytical features needed to define the type of message to be sent to the cluster for each customer. You have homogeneous clusters that will allow you to respond to your business issue. Of particular importance, this segmentation is systematic: the results obtained
do not represent a particular point of view of your data, and is robust or consistent. Two people performing this segmentation using the method would obtain the same results.

6.3 Introduction to Sample Files

This guide is accompanied by the following sample data files:

- A data file *Census01.csv*.
- The corresponding description file *desc_census.csv*.

These files allow you to evaluate Automated Analytics features and take your first steps in using it.

*Census01.csv* is the sample data file that you will use to follow the scenarios for Regression/Classification and Segmentation/Clustering. This file is an excerpt from the American Census Bureau database, completed in 1994 by Barry Becker.

**Note**

For more information about the American Census Bureau, see [http://www.census.gov](http://www.census.gov).

This file presents the data on 48,842 individual Americans, of at least 17 years of age. Each individual is characterized by 15 data items. These data, or variables, are described in the following table.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Example of Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>Age of individuals</td>
<td>Any numerical value greater than 17</td>
</tr>
<tr>
<td>workclass</td>
<td>Employer category of individuals</td>
<td>Private, Self-employed-not-inc</td>
</tr>
<tr>
<td>fnlwgt</td>
<td>Weight variable, allowing each individual to represent a certain percentage of the population</td>
<td>Any numerical value, such as &quot;0&quot;, &quot;2341&quot; or 205019&quot;.</td>
</tr>
<tr>
<td>education</td>
<td>Level of study, represented by a school-level, or by the title of the degree earned</td>
<td>11th, Bachelors</td>
</tr>
<tr>
<td>education-num</td>
<td>Number of years of study, represented by a numerical value</td>
<td>A numerical value between 1 and 16</td>
</tr>
<tr>
<td>marital-status</td>
<td>Marital status</td>
<td>Divorced, Never-married</td>
</tr>
<tr>
<td>occupation</td>
<td>Job classification</td>
<td>Sales, Handlers-cleaners</td>
</tr>
<tr>
<td>relationship</td>
<td>Position in family</td>
<td>Husband, Wife</td>
</tr>
<tr>
<td>race</td>
<td>Ethnicity</td>
<td>White, Black</td>
</tr>
<tr>
<td>Variable</td>
<td>Description</td>
<td>Example of Values</td>
</tr>
<tr>
<td>--------------</td>
<td>----------------------------------</td>
<td>-------------------------------------------------------</td>
</tr>
</tbody>
</table>
| sex          | Gender                           | • Male  
• Female                                          |
| capital-gain | Annual capital gains             | Any numerical value                                    |
| capital-loss | Annual capital losses            | Any numerical value                                    |
| native country| Country of origin                | • United States  
• France                                              |
| class        | Variable indicating whether or not the salary of the individual is greater or less than $50,000 | • "1" if the individual has a salary of greater than $50,000  
• "0" if the individual has a salary of less than $50,000 |

Note

In order to avoid complicating the application scenarios, the variable fnlwgt is used as a regular explanatory variable in these scenarios, and not as a weight variable.

6.4 SAP BusinessObjects Predictive Analytics

To accomplish the scenario, you will use the Automated Analytics toolset of SAP BusinessObjects Predictive Analytics.

6.4.1 To Start SAP BusinessObjects Predictive Analytics

1. Select **Start** > **Programs** > **SAP Business Intelligence** > **SAP BusinessObjects Predictive Analytics Desktop** > **SAP BusinessObjects Predictive Analytics**.

   The SAP BusinessObjects Predictive Analytics start panel appears.

2. Click the feature you want to use.

6.4.2 Editing the Options

- To edit the application options:

  1. In the **File** menu, click **Preferences...**.

     The window **Edit Options...** appears.

  2. The following options can be modified for Segmentation/Clustering:
<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Category</td>
<td>Options</td>
</tr>
</tbody>
</table>
| **General** | ○ Country  
○ Language  
○ Message Level  
○ Log Maximum Size  
○ Message Level for Strange Values  
○ Display the Parameter Tree  
○ Number of Store in the History  
○ Always Exit without Prompt  
○ Include Test in Default Cutting Strategy |
| **Stores** | ○ Default Store for Apply-in Data Set  
○ Default Store for Apply-out Data Set  
○ Default Store to Save Models |
| **Metadata Repository** | ○ Enable Single Metadata Repository  
○ Edit Variable Pool Content |
| **Graphic** | ○ Profit Curve Points  
○ Bar Count Displayed  
○ No InfiniteInsight® Look and Feel  
○ Display 3D Chart  
○ Disable Double Buffering  
○ Optimize for Remote Display  
○ Remember Size and Position when Leaving |
| **Report** | ○ Number of Variables of Interest  
○ Active Style Sheet  
○ Customize Style Sheets |

**Related Information**

*Customizing Style Sheets [page 61]*

### 6.5 Creating a Model Using Modeler

Data modeling with the application is subdivided into four broadly defined stages:

- Defining the Modeling Parameters
- Generation and Validation of the Model
- Analysis and Understanding of the Analytical Results
- Using a Generated Model
6.5.1 Step 1 - Defining the Modeling Parameters

To respond to your business issue, you want to:

- Break down the sample of 50,000 prospects who responded to the test phase of your marketing campaign into homogeneous groups.
- Describe each of these groups and provide customized communication for each of these different groups.

The Segmentation/Clustering feature allows you to create descriptive models.

The first step in the modeling process consists of defining the modeling parameters:

- Select a data source to be used as a training dataset.
- Describe the selected dataset.
- Select the variables.
- Select the explanatory variables.
- Check the Modeling Parameters.
- Define the number of clusters. This step is optional.

Related Information

Describing the Data Selected [page 167]

6.5.1.1 Selecting a Data Source

Use the file Census01.csv as a training dataset.

This file represents the sample that you had extracted from your database and used for the test phase of your direct marketing campaign. As specified in your test plan, this file contains data concerning 50,000 prospects, for whom you now know the behavior with respect to the new financial product:

- 25% of the prospects showed themselves to be clearly interested. They chose to accept an invitation for a meeting with one of your sales channel agents,
- 75% of the prospects declined your invitation.

In this file, you created a new variable Class, which corresponds to the reaction of prospects contacted during the test. You assigned:

- The value "1" to those prospects who responded positively to your invitation,
- The value "0" to those prospects who responded negatively to your invitation.

To select a data source:

1. On the screen Select a Data Source, select the data source format to be used (Text files, Database, ...).
2. Click the Browse button.

   In the Data Selection dialog, browse to the <Installation Path>/Samples folder.
Note

Depending on your environment, the Samples folder may or may not appear directly at the root of the list of folders. If you selected the default settings during the installation process, you will find the Samples folder located in C:Program Files/SAP BusinessObjects Predictive Analytics/Desktop <version number>/Automated/

3. Double-click the Samples folder, then the Census folder.
4. Select the file Census01.csv, then click OK.
   The name of the file appears in the Data Set field.
5. Click Next.
   The screen Data Description appears. Now you are ready to describe the data selected.

6.5.1.2 Describing the Data Selected

For this scenario:

- Select Text Files as the Data Type.
- Use the file Desc_Census01.csv as the description file for the Census01.csv data file.

To Select a Description File:

1. On the screen Data Description, click the button Open Description.
   The Load a Description window opens.
2. Select the type of your description file.
3. In the Folder field, select the folder where the description file is located with the Browse button.
   The folder selected by default is the same as the one you selected on the screen Select a Data Source.
4. In the File Name field, select the file containing the dataset description with the Browse button.

Caution

When the space used for model training contains a physical variable named KxIndex, it is not possible to use a description file without any key for the described space.

When the space used for model training does not contain a physical variable named KxIndex, it is not possible to use a description file including a description about a KxIndex variable since it does not exist in current space.

5. Click OK.
   The window Load a Description closes and the description is displayed on the screen Data Description.
6. Click Next.
6.5.1.2.1 Why Describe the Data Selected?

In order for the application to interpret and analyze your data, the data must be described. To put it another way, the description file must specify the nature of each variable, determining the storage format and type:

- **Storage format**: number (`number`), character string (`string`), date and time (`datetime`) or date (`date`).

**Note**

When a variable is declared as `date` or `datetime`, the date coder feature automatically extracts date information from this variable such as the day of the month, the year, the quarter and so on. Additional variables containing this information are created during the model generation and are used as input variables for the model.

The date coder feature is disabled for Time Series.

- **Type**: `continuous`, `nominal`, `ordinal` or `textual`.

For more information about data description, see the related topics.

6.5.1.2.2 How to Describe Selected Data

To describe your data, you can:

- Either use an existing description file, that is, taken from your information system or a previously created description file from the application features,
- Or create a description file using the Analyze option from the modeling assistant. In this case, you must validate the description file obtained. You can save this file for later use.

**Caution**

The description file obtained using the Analyze option results from the analysis of the first 100 lines of the initial data file. In order to avoid all bias, we encourage you to randomly sort your dataset outside the application before performing this analysis.

6.5.1.2.3 Viewing the Data

To help you validate the description when using the Analyze option, you can display the first hundred lines of your data set.

1. Click the button **View Data**.
   A new window opens displaying the data set top lines:
2. In the field **First Row Index**, enter the number of the first row you want to display.
3. In the field **Last Row Index**, enter the number of the last row you want to display.
4. Click the **Refresh** button to see the selected rows.
6.5.1.2.4  A Comment about Database Keys

For data and performance management purposes, the data set to be analyzed must contain a variable that serves as a key variable. Two cases should be considered:

- If the initial data set does not contain a key variable, a variable index $K_x Index$ is automatically generated by the application. This will correspond to the row number of the processed data.
- If the file contains one or more key variables, they are not recognized automatically. You must specify them manually in the data description.

To Specify that a Variable is a Key

1. In the Key column, click the box corresponding to the row of the key variable.
2. Type in the value "1" to define this as a key variable.

6.5.1.2.5  Defining a Variable Structure

There are three ways to define a variable structure:

- by first extracting the categories from the variable statistics, then editing or validating the suggested structure.
- by importing the structure from an existing model.
- by building a new structure from scratch.

The option Optimal Grouping allows you to let Data Encoding group together the categories groups defined in the variable structure if they bring the same information.

The last column of the description table indicates the state of the structure of each variable. The following table lists the possible states of a variable structure.

<table>
<thead>
<tr>
<th>Icon</th>
<th>State</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Icon]</td>
<td>undefined</td>
<td>Data Encoding will automatically determine the categories grouping depending on their interaction with the target variable.</td>
</tr>
<tr>
<td>![Icon]</td>
<td>non-editable</td>
<td>The structure for an ordinal string variable cannot be modified.</td>
</tr>
<tr>
<td>![Icon]</td>
<td>defined by extraction from the variable statistics</td>
<td>The user must open and validate the variable structure.</td>
</tr>
<tr>
<td>![Icon]</td>
<td>defined by the user or imported from an existing model</td>
<td></td>
</tr>
</tbody>
</table>

Note

A translation of the variable categories has no influence on the variable structure, which has to be set according to the original values of the variable.
Related Information

To Extract a Variable Structure [page 70]
To Import the Variable Structure from a Model [page 71]
To Add a Variable to the List of Variables [page 71]
To Remove a Variable from the List of Variables [page 71]
To View a Variable Structure Defined in the Loaded Model [page 72]
To Create or Modify a Variable Structure [page 72]

6.5.1.3 Filtering the Dataset

In order to accelerate the learn process and to optimize the resulting model, you can apply a filter to your dataset. For more information, see the related topic.

Related Information

Filtering the Data Set [page 76]

6.5.1.4 Translating the Variable Categories

You can translate the categories of a nominal variable, save the translation or load an existing translation. This translation has no influence on the variable structure, which has to be set according to the original values of the variable. For more information, see the related topic.

Related Information

Translating the Variable Categories [page 79]

6.5.1.5 Selecting Variables

Once the training dataset and its description have been entered, you must select different variables:

- One or more targets variables.

  The Segmentation/Clustering feature is capable of segmenting a dataset independently, that is, it does not require that a target variable be selected. However, even though this is not required, we strongly recommend selecting a target variable. For the process of segmenting, a dataset gains maximum meaning
only when it is accomplished in relation to a domain-specific business issue, expressed in the form of a target variable.

- Possibly a weight variable.
- The explanatory variables.

6.5.1.5.1 Target Variables

For this scenario, select the variable Class as your target variable, that is, the variable that indicates the probability of an individual responding in a positive or negative manner to your campaign.

1. On the screen Selecting Variables, in the section Explanatory variables selected (left hand side), select the variables you want to use as Target Variables.

   **Note**
   
   On the screen Selecting Variables, variables are presented in the same order as that in which they appear in the table of data. To sort them alphabetically, select the option Alphabetic sort, presented beneath each of the variables list.

2. Click the button > located on the left of the screen section Target Variables (upper right hand side).

   The variable moves to the screen section Target Variables.

   You can also select a variable in the screen section Target Variables and click the button < to move the variables back to the screen section Explanatory variables selected.

6.5.1.5.2 Weight Variable

Selecting a Weight Variable enables to set the Weight Quantum option available in the Advanced Model Parameters.

For this scenario, do not select a weight variable.

1. On the screen Selecting Variables, in the section Explanatory variables selected (left hand side), select the variables you want to use as Weight Variables.

   **Note**
   
   On the screen Selecting Variables, variables are presented in the same order as that in which they appear in the table of data. To sort them alphabetically, select the option Alphabetic sort, presented beneath each of the variables list.

2. Click the button > located on the left of the screen section Weight Variables (middle right hand side).

   The variable moves to the screen section Weight Variables.

   You can also select a variable in the screen section Weight Variables and click the button < to move the variables back to the screen section Explanatory variables selected.
6.5.1.5.3 Explanatory Variables

By default, and with the exception of key variables, all variables contained in your dataset are taken into consideration for generation of the model. You may exclude some of these variables.

The decision whether to include or exclude a variable for generation of your segmentation model depends upon domain-specific considerations. Your domain-specific knowledge allows you to determine which variables are the most useful for description of the clusters or homogeneous groups. A regression model generated using Regression/Classification would also be used as a tool to determine the variables with the greatest explanatory power for a given phenomenon.

For this Scenario:

- Exclude the variable KxIndex, as this is a key variable. Since the initial data set does not contain a key variable, the application generated KxIndex automatically.
- Retain all the other variables.

1. To exclude some variables from data analysis, on the screen Selecting Variables, in the section Explanatory Variables Selected (left hand side), select the variable to be excluded.
   
   ! Note
   
   On the screen Selecting Variables, variables are presented in the same order as that in which they appear in the table of data. To sort them alphabetically, select the option Alphabetic sort, presented beneath each of the variables list.

2. Click the button > located on the left of the screen section Excluded Variables (lower right hand side). The variable moves to the screen section Excluded Variables. Also, select a variable in the screen section Excluded Variables and click the button < to move the variables back to the screen section Explanatory Variables Selected.

3. Click Next.

The screen Parameters of the Model appears.

6.5.1.6 Checking Modeling Parameters

The screen Summary of Modeling Parameters allows you to check the modeling parameters just before generating the model.

! Note

The screen Summary of Modeling Parameters contains an Advanced button. By clicking this button, you access the screen Advanced Model Parameters. For more information about these parameters, Setting Up the Advanced Options.

The Model Name is filled automatically. It corresponds to the name of the target variable (class for this scenario), followed by the underscore sign ("_"), and the name of the data source, minus its file extension (Census01 for this scenario).

Before generating the model, you can define the number of clusters that you want to obtain. These fields allow you to specify how many clusters will be generated by the model. By default the number of clusters is set to 10.
The higher the number of segments, the lower the robustness (KR). The lower the number of segments, the lower the information (KI). One should generally start with the default number and then go further with more or fewer clusters based on the results.

- For supervised segmentation (with a target), choose the best number of segments, for example, [5;10] means that you are requesting 5 to 10 clusters. The application computes the "best number of clusters" using the metric KI+KR. For instance, you may have 7 clusters.
- For unsupervised segmentation (without target), the application chooses the minimum number of clusters, for instance [10;10] which means that you requested 10 clusters.

Choosing to Calculate SQL Expressions allows you to see, in the model debriefing, the SQL Expressions used to generate each cluster.

**Note**

When you activate the option Calculate SQL Expressions, the application generates an additional cluster that contains the unassigned records. For more details on SQL expressions and unassigned records, see Difference Between Standard Cross Statistics and SQL Expressions.

For this scenario, keep the default settings.

### Related Information

Setting Up the Advanced Options [page 173]
Difference Between Standard Cross Statistics and SQL Expressions [page 201]

#### 6.5.1.6.1 Setting Up the Advanced Options

The Advanced Model Parameters panel provides you with several options.

You may calculate the cross-statistics for the model to be generated, define the target key value, choose the distance computing option or choose the encoding strategy.

### Calculating the Cross Statistics

The Set Target Keys value option lists the target variables selected in the Selecting Variables screen and allows you to choose their key value.

To Define the target key value in the Target Key field, enter the key value of the target variable.

### Choosing the Distance Computing Method

The Distance list allows you to specify the distance used to compare input data encoded by the data encoder.
To choose the distance computing method, in the Distance drop-down list, select among these options:

- **Chessboard**: maximum of absolute differences between coordinates (LInf).
- **Euclidean**: square root of sum of square differences between coordinates (L2).
- **City Block**: sum of absolute differences between coordinates (L1).
- **System Determined** (default value): Lets the system determine the best distance to be used according to the model build settings.

**Note**

The current policy is to use LInf either in unsupervised mode or when the clusters SQL expressions have been asked for, and L2 otherwise.

### 6.5.1.6.1.1 Encoding Strategy

The **Encoding Strategy** option refers to the kind of encoding the segmentation engine is expecting from the data encoder of Automated Analytics.

- **To Choose an Encoding Strategy:**
  - Choose among the following options from the drop-down list:

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>System Determined</td>
<td>Lets the system select the best encoding according to the model parameters. The <strong>Target Mean</strong> encoding is used for supervised models. Otherwise, variables are encoded using the <strong>Unsupervised</strong> scheme.</td>
</tr>
<tr>
<td>Target Mean</td>
<td><strong>Default value for supervised clustering</strong></td>
</tr>
<tr>
<td></td>
<td>Each value of a continuous input variable is replaced by the mean of the target for the segment the value belongs to.</td>
</tr>
<tr>
<td></td>
<td>Each category of a nominal input variable is replaced by the mean of the target for this category.</td>
</tr>
<tr>
<td></td>
<td>In case of a nominal target variable, the mean of the target corresponds to the percentage of positive cases of the target variable for the input variable category.</td>
</tr>
<tr>
<td>Uniform</td>
<td>Each variable segment is encoded in the range ([-1;+1]) so that the distribution of the variables is uniform.</td>
</tr>
<tr>
<td>Unsupervised</td>
<td><strong>Default value for unsupervised clustering</strong></td>
</tr>
<tr>
<td></td>
<td>A target free strategy. Only segment frequency is used to encode variables.</td>
</tr>
</tbody>
</table>

The following options will only be displayed when all variables are continuous:

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural</td>
<td>This option does not transform the input data.</td>
</tr>
</tbody>
</table>
### 6.5.1.6.2 Activating the Autosave Option

The **Model Autosave** panel allows you to activate the option that will automatically save the model at the end of the generation process and to set the parameters needed when saving the model.

To activate the option, proceed as follows:

1. In the **Summary of Modeling Parameters** panel, click the **Autosave** button. The **Model Autosave** panel is displayed.
2. Check **Enable Model Autosave**.
3. Set the parameters listed in the following table.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Model Name</strong></td>
<td>This field allows you to associate a name with the model. This name will then appear in the list of models to be offered when you open an existing model.</td>
</tr>
<tr>
<td><strong>Description</strong></td>
<td>This field allows you to enter the information you want, such as the name of the training data set used, the polynomial degree or the performance indicators obtained. This information could be useful to you later for identifying your model. Note that this description will be used instead of the one entered in the <strong>Summary of Modeling Parameters</strong> panel.</td>
</tr>
</tbody>
</table>
| **Data Type** | This list allows you to select the type of storage in which you want to save your model. The following options are available:  
  ○ **Text Files**, to save the model in a text file.  
  ○ **Data Base**, to save the model in a database.  
  ○ **Flat Memory**, to save the model in the active memory. |
| **Folder** | Depending upon which option you selected, this field allows you to specify the ODBC source, the memory store or the folder in which you want to save the model. |
| **File/Table** | This field allows you to enter the name of the file or table that is to contain the model. When saving the model as a text file, you must enter one of the following format extensions .txt (text file in which the data is separated by tabs) or .csv (text file in which the data is separated by commas). |

4. Click **OK**.
6.5.2 Step 2 - Generating and Validating the Model

Once the modeling parameters are defined, you can generate the model. Then you must validate its performance using the predictive power (KI) and the prediction confidence (KR):

- If the model is sufficiently powerful, you can analyze the responses that it provides in relation to your business issue (see Step 3 - Analyzing and Understanding the Model Generated), and then apply it to new datasets (see Step 4 - Using the Model).
- Otherwise, you can modify the modeling parameters in such a way that they are better suited to your dataset and your business issue, and then generate new, more powerful models.

Related Information

Step 3 - Analyzing and Understanding the Model Generated [page 179]
Step 4 - Using the Model [page 205]

6.5.2.1 Generating the Model

To generate the model:

1. On the Advanced Model Parameters screen, click Generate.
   The screen Training the Model appears. The model is being generated. A progress bar allows you to follow the process.
2. If the Autosave option has been activated in the panel Summary of Modeling Parameters, a message is displayed at the end of the learning process confirming that the model has been saved.
3. Click Close.
4. Once the model has been generated, click Next to go to panel Using the Model.

6.5.2.2 Following the Progress of the Generation Process

There are two ways for you to follow the progress of the generation process:

- The Progress Bar displays the progression for each step of the process. It is the screen displayed by default.
- The Detailed Log displays the details of each step of the process.

To display the Progression Bar

Click View Type and select (Progress).
The progression bar screen appears.

**To Display the Detailed Log**

Click **View Type** and select the (Log) button.

The detailed log displays the details of each step of the process.

**To Stop the Learning Process**

1. Click the (Stop Current Task) button.
2. Click the **Previous** button.
   The screen **Summary of Modeling Parameters** appears.
3. Go back to the section on checking modeling parameters.

### 6.5.2.3 Validating the Model

Once the model has been generated, you must verify its validity by examining the performance indicators:

- The predictive power allows you to evaluate the explanatory power of the model, that is, its capacity to explain the target variable when applied to the training dataset. A perfect model would possess a predictive power equal to 1 and a completely random model would possess a predictive power equal to 0.
- The prediction confidence defines the degree of robustness of the model, that is, its capacity to achieve the same explanatory power when applied to a new dataset. In other words, the degree of robustness corresponds to the predictive power of the model applied to an application dataset.

To see how the predictive power and the prediction confidence are calculated, see Predictive Power, Prediction Confidence and Model Graphs.

Besides the Predictive Power (KI) and the Prediction Confidence (KR), the application also provides you with two commonly known indicators:

- the classification rate, in case of a classification model,
- the Pearson Square Correlation coefficient (named R2 in the application), in case of a regression model.

Both indicators can be used to compare the application results with results obtained through other data mining tools.

**Note**

Validation of the model is a critically important phase in the overall process of Data Mining. Always be sure to assign significant importance to the values obtained for the predictive power and the prediction confidence of a model.
To validate a segmentation model, you can also observe the value of the indicators “frequency” and “target mean” for each of the identified clusters. Specifically, the most interesting clusters of the segmentation model will possess an elevated “frequency” and a “target mean” that deviates from the “target mean” of the entire dataset. Note that a segmentation model with a low predictive power may conceal precisely this type of cluster.

To find out how the frequency and target mean for a cluster are calculated, see Understanding the Detailed Description of Clusters.

For this scenario, the model generated possesses:
- A predictive power equal to 0.703.
- A prediction confidence equal to 0.987.

The model performs sufficiently well. You do not need to generate another.

To Validate the Model Generated

1. Verify the Predictive Power (KI) and the Prediction Confidence (KR) of the model.

   **Note**
   
   As a general note, other indicators are provided in addition to the predictive power and the prediction confidence during the generation of the model. For example, you could view the Learning Time required to generate the model and information on the targets.

2. To verify the indicators in the detailed log, click \( \text{Log} \).
3. You can then display the screen Using the Model.
   - If the performance of the model satisfies you, go to Step 3 - Analyzing and Understanding the Model Generated.
   - Otherwise, go to the procedure To Generate a New Model.

To Generate a New Model

You have two options. On the screen Training the Model, you can:
- Either click the Previous button to return to the modeling parameters defined initially. Then you can modify the parameters one by one.
- Or click the Cancel button to return to the main screen of the modeling assistant. Then you must redefine all the modeling parameters.

Related Information

Predictive Power, Prediction Confidence and Model Graphs [page 185]
Step 3 - Analyzing and Understanding the Model Generated [page 179]
6.5.3 Step 3 - Analyzing and Understanding the Model Generated

The suite of plotting tools within the application allows you to analyze and understand the model generated:

- The performance of the model with respect to a hypothetical perfect model and a random type of model,
- The characteristics of each of the clusters,
- The significance of the various categories of each variable with respect to the target variable (cross statistics).

6.5.3.1 Presentation of the User Menu

Once the model has been generated, Click Next. The screen Using the Model appears.

The screen Using the Model presents the various options for using a model, that allow you to:

- Display the information relating to the model just generated or opened (Display section), referring to the model curve plots, plotting of clusters, contributions by variables and the profiles of variables of each cluster.
- Apply the model just generated or opened to new data (Run section).
- Save the model, or generate the source code (Save/Export section).

6.5.3.2 Model Overview

The screen Model Overview displays the same information as the training summary.

### Overview

<table>
<thead>
<tr>
<th>Model</th>
<th>name of the model, created from the target variable name and the dataset name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset</td>
<td>name of the dataset</td>
</tr>
<tr>
<td>Initial Number of Variables</td>
<td>number of variables in the dataset</td>
</tr>
<tr>
<td>Number of Selected Variables</td>
<td>number of explanatory variables used to build the model</td>
</tr>
<tr>
<td>Number of Records:</td>
<td>number of records in the dataset</td>
</tr>
<tr>
<td>Building Date</td>
<td>date and time when the model was built</td>
</tr>
<tr>
<td>Learning Time</td>
<td>total learning time</td>
</tr>
<tr>
<td>Engine name</td>
<td>name of the feature used to build the model: Kxen.KMeans for a segmentation</td>
</tr>
<tr>
<td>Minimum / Maximum Requested Number of Clusters</td>
<td>number of clusters that have been asked for by the user</td>
</tr>
</tbody>
</table>
SQL Expressions indicates if the SQL expressions for the clusters definitions have been calculated (Enabled ) or not (Disabled )

Nominal Target Variables

For each nominal target:

<table>
<thead>
<tr>
<th><code>&lt;TargetVariableName&gt;</code></th>
<th>name of the target variable for which the statistics are displayed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target Key</td>
<td>wanted target value</td>
</tr>
<tr>
<td><code>&lt;NonTargetCategory&gt;</code></td>
<td>Frequency</td>
</tr>
<tr>
<td></td>
<td>frequency in percentage of the non-target value in the entire dataset</td>
</tr>
<tr>
<td><code>&lt;TargetCategory&gt;</code></td>
<td>Frequency</td>
</tr>
<tr>
<td></td>
<td>frequency in percentage of the wanted target value in the entire dataset</td>
</tr>
</tbody>
</table>

Continuous Targets (Number)

For each continuous target:

<table>
<thead>
<tr>
<th><code>&lt;TargetVariableName&gt;</code></th>
<th>name of the target variable for which the statistics are displayed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>minimum value for the target</td>
</tr>
<tr>
<td>Max</td>
<td>maximum value for the target</td>
</tr>
<tr>
<td>Mean</td>
<td>mean of the target</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>mean of the distance between the target values and the Mean</td>
</tr>
</tbody>
</table>

Performance Indicators

For each target variable:

<table>
<thead>
<tr>
<th>Predictive Power (KI)</th>
<th>For more information on the predictive power, see section Performance indicators (on page 39 ).</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prediction Confidence (KR)</td>
<td>For more information on the prediction confidence, see section Performance indicators (on page 39 )</td>
</tr>
</tbody>
</table>
### Cluster Counts

For each target variable:

<table>
<thead>
<tr>
<th>Cluster Counts</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial Number of Clusters</td>
<td>number of clusters that have been asked for by the user</td>
</tr>
<tr>
<td>Final Number of Clusters</td>
<td>number of clusters found by the model</td>
</tr>
</tbody>
</table>

### 6.5.3.2.1 Model Overview Options

#### To Copy the Model Overview

Click the (Copy) button.

The application copies the HTML code of the screen. You can paste into a word processing or spreadsheet program, a text editor, ...

#### To Save the Model Overview

Click the (Save) button situated under the title.

The file is saved in HTML format.

#### To Print the Model Overview

1. Click the (Print) button situated under the title.
   A dialog box appears, allowing you to select the printer to use.
2. Select the printer to use and set other print properties if need be.
3. Click OK.
   The report is printed.

#### To Export to PowerPoint

Click the (Export to PowerPoint) button.
6.5.3.3 Model Graphs

The model graphs allow you to:

- View the realizable profit that pertains to your business issue using the model generated.
- Compare the performance of the model generated with that of a random type model and that of a hypothetical perfect model.

On the plot, for each type of model, the curves represent the realizable profit (Y axis, or ordinate) as a function of the ratio of the observations correctly selected as targets relative to the entire initial dataset (X axis, or abscissa).

Displaying the Model Graphs

1. On the screen Using the Model, click the Model Graphs option. The model graph appears.

![Model Graphs Diagram]

The default parameters display the profit curves corresponding to the Validation sub-set (blue line), the hypothetical perfect model (Wizard: green line) and a random model (Random: red line). The default setting for the type of profit parameter is Detected profit, and the values of the abscissa are provided in the form of a percentage of the entire dataset.

2. When there is more than one target, select the target for which you want to see the curves in the Models list.

   \[ \text{Note} \]

   One model corresponds to each variable. The name of each model is built from the kC_prefix and the model target name.

3. Select the viewing options that interest you. For more information about viewing options.
6.5.3.3.1 Plot Options

To Display the Graphs for the Estimation, Validation and Test Sub-sets

Click Dataset and select one of the following options that allow you to switch between:

- the graph for the Validation sub-set: ![Graph]
- the graphs for all the sub-sets: ![Graph]

To Change the View Type

Click View Type and select the desired option.

To Copy the Model Graph

Click the (Copy) button and select the desired option.

The application copies the parameters of the plot. You can paste it into a spreadsheet program (such as Excel) and use it to generate a graph.

To Save the Model Graph

1. Click the (Save) button.
   A dialog box appears, allowing you to select the file properties.
2. Type a name for your file.
3. Select the destination folder.
4. Click OK.
   The plot is saved as a PNG formatted image.

To Print the Model Graph

1. Click the (Print) button situated under the title.
   A dialog box appears, allowing you to select the printer to use.
2. Select the printer to use and set other print properties if need be.
3. Click OK. The report is printed.

To Export the Model Graph to Microsoft Excel

Click the (Export to Excel) button situated under the title. An Excel sheet opens containing the model graph you are currently viewing along with its data.

To Open the Current Graph in a New Window

Click the (Pin View) button. The current graph is displayed in a new window.

6.5.3.3.2 Understanding the Model Graphs

6.5.3.3.2.1 Understanding Graphs for a Model with a Nominal Target

The following figure represents the model graph produced using the default parameters.

On the plot, the curves for each type of model represent the profit that may be realized (Y axis), that is, the percentage of observations that belong to the target variable, in relation to the number of observations selected from the entire initial dataset (X axis). On the X axis, the observations are sorted in terms of
decreasing "score", that is, the decreasing probability that they belong to the target category of the target variable.

In the application scenario, the model curves represent the ratio of prospects likely to respond in a positive manner to your marketing campaign, relative to the entire set of prospects contained in your database.

Detected profit is the default setting for type of profit. Using this type of profit:

- The value "0" is assigned to observations that do not belong to the target category of the target variable,
- The value "1/(frequency of the target variable in the dataset)" is assigned to observations that do belong to the target category of the target variable.

The following table describes the three curves represented on the plot created using the default parameters.

<table>
<thead>
<tr>
<th>The curve...</th>
<th>Represents...</th>
<th>For instance, by selecting...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wizard (green curve, at the top)</td>
<td>The profit that may be achieved using the hypothetical perfect model that allows one to know with absolute confidence the value of the target variable for each observation of the dataset</td>
<td>25% of the observations from your entire dataset with the help of a perfect model. 100% of observations belonging to the target category of the target variable are selected. Thus maximum profit is achieved. Note that these 25% correspond to the proportion of prospects who responded in a positive manner to your marketing campaign, during your test phase. For these prospects, the value of the target variable, or profit, is equal to 1.</td>
</tr>
<tr>
<td>Validation (blue curve, in the middle)</td>
<td>The profit that may be achieved using the model generated by Modeler - Regression/Classification that allows one to perform the best possible prediction of the value of the target variable for each observation of the dataset</td>
<td>25% of the observations from your initial dataset with the help of the model generated, 66.9% of the observations belonging to the target category of the target variable are selected.</td>
</tr>
<tr>
<td>Random (red curve, at the bottom)</td>
<td>The profit that may be achieved using a random model that does not allow one to know even a single value of the target variable for each observation of the dataset</td>
<td>25% of the initial dataset using a random model, 25% belonging to the target category of the target variable are selected.</td>
</tr>
</tbody>
</table>

6.5.3.3.3 Predictive Power, Prediction Confidence and Model Graphs

On the model graph plot:

- Of the estimation data set (default plot), the predictive power corresponds to “the area found between the curve of the model generated and that of the random model” divided by “the area found between the curve of the perfect model and that of the random model”. As the curve of the generated model approaches the curve of the perfect model, the value of the predictive power approaches 1.
- Of the estimation, validation and test data sets (select the corresponding option from the list Data set, located below the plot), the prediction confidence corresponds to one minus “the area found between the
curve of the estimation data set and that of the validation data set" divided by "the area found between the curve of the perfect model and that of the random model".

The following graph displays the predictive power and the prediction confidence:

![Graph showing predictive power and prediction confidence](Image)

6.5.3.4 Category Significance

Definition

The Significance of Categories plot illustrates the relative significance of the different categories of a given variable with respect to the target variable.

Displaying the Significance of Categories Plot

1. On the screen Using the Model, click Category Significance. The plot Category Significance appears.
2. In the Variables list located above the plot, select the variable for which you want to display the categories. If your dataset contains date or datetime variables, automatically generated variables can appear in the Variables list. For more information, refer to section on date and date-time variables in the topic on variable storage formats.

**Note**
- You can display the relative significance of the categories of a variable directly from the plot Contributions by Variables. On the plot Contributions by Variables, double-click the bar of the variable which interests you.
- In case no user structure has been defined for a continuous variable, the plot category significance displays the categories created automatically using the band count parameter. The number of categories displayed corresponds to the value of the band count parameter. For more information about configuring this parameter, please refer to the section Band Count for Continuous Variables.

**Plot Options**

**To Switch Between "Validation Data Set" and "All Data Sets" Plots**

1. Click Data Sets and select the (All Data Sets) button to display all datasets.
The plot displaying all datasets appears.

2. Click *Data Sets* and select the *(Validation Only)* button to go back to the *Validation Data Set* plot.

**To Switch between Curve and Bar Charts**

1. Click *View Type* and select the button to display the curve chart. The curve plot appears.

2. Click *View Type* and select the button to go back to the bar chart.

**Note**
You can combine the different types of plot. For example, you can display *All Datasets* in a curve chart or the *Validation Data Set* in a bar chart.

**Understanding the Plots of Variables**

For this scenario, select the variable *marital-status*, which is the explanatory variable that contributes the most to the target variable *Class*. This plot presents the effect of the categories of the *marital-status* variable on the target variable. For an explanation of the plot, see the topic on variable categories and profit.

**Related Information**

Variable Storage Formats [page 30]
Band Count for Continuous Variables [page 75]
Variable Categories and Profit [page 188]

**6.5.3.4.1 Variable Categories and Profit**

The *Category Significance* plot shows the relative significance of the different categories of a given variable with respect to the target variable.

Each bar on the plot shows the amount of influence that category has on the target category (or hoped-for value) of the target variable. The bars are ordered by their amount of influence. In this example, the *Bar Orientation* option shows the bars vertically, so the bar the furthest to the left represents the category with the greatest positive effect, and the bar the furthest to the right, the category with the least effect.
If the **Bar Orientation** option is set to show the bars horizontally, the highest bar on the plot represents the category with the greatest positive effect. In other words, the higher a category appears on the plot, the more representative that category is of the target category of the target variable.
The length and direction of a bar correspond to the profit contributed by that category, in other words, whether the category has more or fewer observations belonging to the target category. A positive bar (Influence on Target greater than 0.0) indicates that the category contains more observations belonging to the target category than the mean (calculated on the entire dataset). A negative bar (Influence on Target less than 0.0) indicates that the category contains fewer observations belonging to the target category than the mean.

**Note**

You can display the profit curve for the selected variable by clicking the (Display Profit Curve) button located in the tool bar under the title.

The importance of a category depends on both its difference to the target category mean and the number of represented cases. High importance can result from any of the following:

- A high discrepancy between the category and the mean of the target category of the target variable
- A minor discrepancy combined with a large number of records in the category
- A combination of both

Use the Variables pull down menu to select and graph any of the variables in the model. Use the tool bar located under the title to copy the coordinates to the clipboard, print the plot, or save it in PNG format. The values are normalized and their sum always equals to "0". Depending on the chosen profit strategy, or on the continuous target variables value type, you can obtain all positive importance or negative and positive importance.
### 6.5.3.4.1.1 Axes

The X-axis shows the influence of the variable categories on the target. The significance of the different numbers on the X-axis are detailed in the following table:

<table>
<thead>
<tr>
<th>Number on the X-axis</th>
<th>Indicates that the category has</th>
</tr>
</thead>
<tbody>
<tr>
<td>positive number</td>
<td>a positive influence on the target</td>
</tr>
<tr>
<td>0</td>
<td>no influence on the target (the behavior is the same as the average behavior of the whole population)</td>
</tr>
<tr>
<td>negative number</td>
<td>a negative influence on the target</td>
</tr>
</tbody>
</table>

The Y-axis displays the variable categories. Categories sharing the same effect on the target variable are grouped. They appear as follow [Category_a;Category_b;Category_c]. Categories not containing sufficient numbers to provide robust information are grouped in the KxOther category. When a variable is associated with too many missing values, the missing values are grouped in the KxMissing category. Both categories are created automatically by the application.

### 6.5.3.4.1.2 Formulas

Category Importance = \( NP \times BF / NC \) where \( NP \) is the **Normal Profit**, \( BF \) is the **Bin Frequency** and \( NC \) is the **Normalization Constant**.

The calculation of the normalization constant differs by target data type. The calculations for binary and continuous targets are detailed below.

For binary targets, it is calculated as follows:

\[
(Target \ Frequency) \times (1 - [Target \ Frequency]).
\]

It can be approximated for non-pathological continuous targets (that is continuous targets without distribution peak (Dirac)) from:

\[
(Proportion \ above \ Median) \times (1 - Proportion \ above \ Median) = 0.5 \times (1 - 0.5) = 0.25
\]

### 6.5.3.4.2 Grouping Categories

On the plot of details of a variable, categories may appear grouped.

When the option **Optimal Grouping** is enabled, the application groups those categories sharing the same effect on the target variable. In the plot below (example in English), for the variable education, the categories Doctorate and Prof-School are grouped. If the explanatory variable is continuous, the application identifies the points where behavioral changes occur with respect to the target variable and automatically crops the variable into intervals exhibiting homogeneous behavior with respect to the target.

For more information, please see the related topic.
When categories do not contain sufficient numbers to provide robust information, they are grouped in the KxOther category, that is created automatically.

When a variable is associated with too many missing values, the missing values are grouped in the KxMissing category, that is also created automatically.

To understand the value of the categories KxOther and KxMissing, consider the following example. The database of corporate customers of a business contains the variable "web address". This variable contains the Web site address of the corporate customers contained in the database. Some companies have a Web site; others do not. In addition, each Web site address is unique. In this case, the application automatically transforms the "web address" variable into a binary variable with two possible values: KxOther (the firm has a Web site) and KxMissing (the firm does not have a Web site).

**Related Information**

Optimal Grouping for All Variables [page 76]
6.5.3.5  Cluster Summary

The following types of charts can be displayed:

- Bubble Charts: Bubble charts display the clusters by representing the relationship between three variables.
- Bar Charts (Cluster Plots): Bar charts display the three cluster plots that allow you to examine:
  - The proportion of observations of the dataset contained in each cluster (Frequencies plot).
  - The proportion of each cluster relative to the target variable (Target Means and Relative Target Means plots).

<i>Note</i>

An additional cluster is always displayed in the cluster summary. This cluster contains all the data points that are not included in the clusters computed by the product.

6.5.3.5.1  Displaying Bubble Charts

1. On the screen Using the Model, click Cluster Summary.
   The panel Cluster Summary appears.
2. Use the options to define the variables you want to display in the bubble chart.

   The table below lists the available options:

<table>
<thead>
<tr>
<th>The option...</th>
<th>allows you to...</th>
<th>Note that...</th>
</tr>
</thead>
<tbody>
<tr>
<td>select the variable to be used in the X-Axis.</td>
<td>only continuous and nominal numerical variables can be used.</td>
<td></td>
</tr>
<tr>
<td>select the variable to be used in the Y-Axis.</td>
<td>only continuous and nominal numerical variables can be used.</td>
<td></td>
</tr>
<tr>
<td>select the variable to be used for the bubble size.</td>
<td>only the variable Frequency and the target variable can be used.</td>
<td></td>
</tr>
<tr>
<td>display cluster names.</td>
<td>cluster names can be customized in Cluster Profiles.</td>
<td></td>
</tr>
</tbody>
</table>

6.5.3.5.2  Understanding Bubble Charts

Bubble charts allow you to display the clusters representing the relationship between three variables. Thus, a bubble chart can provide 3 pieces of information on each cluster.

In addition, bubble charts provide a graphical representation of the segmentation, enabling you to easily visualize the clusters. For instance, it can be useful during a presentation.
The X-axis, the Y-axis and the bubble size represent one variable each. You can define the variables to use in the chart. Thus, you can create a bubble chart that separates distinctly the clusters from each other, enabling you to identify the clusters of interest for your marketing campaign.

The figure below represents the relationship between the variables class, capital-gain and frequency.

For instance, results show that the customers listed in cluster 3 are earning 8,682.24 dollars per year on average (capital-gain: 8,682.24) and represents 7% (Frequency: 0.07) of the population listed in the dataset. In addition, among these customers, 75% (class: 0.75) responded in a positive manner to the test phase of your marketing campaign.

In comparison, cluster 2 represents the biggest population listed in the dataset, namely 24% of the population (Frequency: 0.24), which is around three times bigger than the population listed in cluster 3. However, the customers listed in this cluster are earning less than the customers listed in cluster 3. They earn on average 38.40 dollars per year (capital-gain: 38.40). In other words, less than 0.5% of what customers listed in cluster 3 earn. Moreover, among the customers listed in cluster 2, only 4% (class: 0.04) responded in a positive manner to the test phase of your marketing campaign.

Consequently, compared to cluster 2, cluster 3 is more interesting because it shows better results to the test phase of your marketing campaign.

6.5.3.5.3 Displaying Cluster Plots

1. On the screen Using the Model, click Cluster Summary.

   The panel Cluster Summary appears.

2. Click the button (View Type) and select Bar Chart.

3. In the Chart Type list, select the type of chart that you want to display.

   Note

   Select the option Descending sort to sort the plot bars in descending order. For instance, on the plot Relative Target Means, the descending sort allows quick examination of the most interesting clusters, that is, those which differ most from the mean behavior of the dataset taken as a whole.
6.5.3.5.4 Understanding Cluster Plots

**The Target Means Plot**

The Target Means plot presents the proportion of observations belonging to the target category of the target variable, present in each cluster.

The figure below presents the Target Means plot obtained during this scenario. The bars have been sorted in descending order.

Among the five clusters, Cluster 6 is the one that has the greatest proportion of observations belonging to the target category of the target variable. In fact, 85.5% of the customers contained in cluster 6 belong to target category 1, target variable Class. In other words, 85.5% of the customers contained in cluster 6 responded in a positive manner to the test phase of your marketing campaign.

Cluster 8 is the cluster with the lowest density of observations belonging to the target category. Less than 1% of the customers contained in this cluster responded positively to the test phase of your marketing campaign.

**The Frequencies Plot**

The Frequencies plot presents the number of observations contained in each cluster relative to the total number of observations contained in the dataset. The figure below presents the Frequencies plot obtained during this scenario. The bars have been sorted in descending order. The Frequencies plot presents the number of observations contained in each cluster relative to the total number of observations contained in the dataset.

The figure below presents the Frequencies plot obtained during this scenario. The bars have been sorted in descending order.
Among the five clusters, Cluster 2 is the one which contains the greatest number of observations, or 25.2% of the total number of customers contained in the entire dataset.

**The Relative Target Means Plot**

Similar to the Target Means plot, the Relative Target Means plot presents the proportion of observations, for each cluster, belonging to the target category of the target variable. The only difference between the two plots is the scale used on the Y axis. On the Relative Target Means plot, the proportion of observations belonging to the target category of the target variable relative to the entire dataset is re-expressed. In other words, the 0 value of the Y axis corresponds to the true percentage of observations belonging to the target category of the target variable in relation to the entire dataset.

The figure below presents the Relative Target Means plot obtained during this scenario. The bars have been sorted in descending order.
Among the ten clusters, Cluster 6 is the cluster that has the highest proportion of observations belonging to the target category of the target variable. Compared to the entire dataset, Cluster 6 contains 61.6% more customers belonging to the target category 1, of the target variable Class.

Cluster 2 contains less than 3.3% of customers belonging to the target category. In other words, Cluster 2 has almost the same customer density belonging to the target category as the dataset taken as a whole.

Cluster 8 is the cluster with the lowest density of observations belonging to the target category. Compared to the entire dataset, Cluster 8 contains 23.5% fewer of customers belonging to the target category. This cluster therefore has a density of customers belonging to the target category lower than that of the dataset.

6.5.3.6 Cluster Profiles

6.5.3.6.1 Cross-Statistics and Variable Profiles

The cluster profiles allow you to view for each cluster:

- the profile of each explanatory variable, with respect to their profile over the entire dataset,
- the SQL expression of the cluster when they have been calculated.

Variable Profile

The Variable Profile indicates the distribution of observations (belonging to a cluster of global dataset) within the categories of each variable. In other words, the profile indicates the proportion of observations contained in each of the categories of that variable.

Example of a Variable Profile

The variable "gender" of a dataset can be distributed as follows:

- 53% of observations belong to the category "male".
- 47% of observations belong to the category "female".

This distribution corresponds to the profile of the variable "gender" over this dataset.

Given a cluster A, taken from this dataset, the same variable "gender" may be distributed as follows:

- 80% of observations belong to the category "male".
- 20% of observations belong to the category "female".

This distribution corresponds to the profile of the variable "gender" over cluster A.

The cluster profiles allow you to view and compare the profiles of the variable "gender" over the dataset and the clusters taken from this dataset.
6.5.3.6.2 Displaying Cluster Profiles

1. On the screen Using the Model, click Cluster Profiles.
   The screen Cluster Profiles appears.

2. In the table, select the cluster for which you want to view the profile.

   i Note
   If only the variable ranges for a specific cluster are displayed, click the black horizontal bar, a moving cursor is displayed. Drag the cursor down to display the list of clusters.

3. Below the table, from the drop-down list associated with the Variable field, select the variable for which you want to see the profile.
   The cross statistics will appear in the form of a plot in the lower part of the screen.

6.5.3.6.3 Understanding Cluster Profiles

The screen Cluster Profiles can be broken down into three parts:

- In the upper part, a table summarizes the information for each cluster. This table allows you to select the cluster for which you want to view cross statistics.

<table>
<thead>
<tr>
<th>Column</th>
<th>Indicates...</th>
<th>For instance...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster Name</td>
<td>The name of the cluster.</td>
<td>1</td>
</tr>
<tr>
<td>Frequencies</td>
<td>The number of observations contained in the cluster relative to the total number of observations contained in the dataset.</td>
<td>The customers contained in cluster 1 represent 4.22% of the total number of customers contained in your entire training dataset.</td>
</tr>
<tr>
<td>% of '1'</td>
<td>The proportion of observations contained in the cluster belonging to the target category of the target variable.</td>
<td>51.17% of the customers contained in cluster 1 belong to the target category of the target variable Class. In other words, 51.17% of the customers contained in this cluster responded in a positive manner to the test phase of your marketing campaign.</td>
</tr>
</tbody>
</table>

- In the middle part, a drop-down list allows you to select the variable for which you want to see the cross statistics. Variables are presented in descending order of the significance of their contribution relative to the target category of the target variable. When a cluster is selected, the variables visible in the drop-down list are sorted according to the difference between their cluster profile and their population profile (the Kullback-Leibler divergence is used to measure this difference). The variable that appears first on the list is the variable exhibiting the greatest difference between its two profiles. This sorted list of variables provides the set of discriminatory variables required to describe a cluster.

- In the lower part, a plot presents either the cross statistics corresponding to the cluster and the variables selected, or, when it has been calculated, the SQL expression defining the cluster.

The figure below presents the screen Cluster Profiles, which appears as the default plot for this scenario. The plot presents the cross statistics for cluster 6.
Cross Statistics Plots

Cross statistics plots contain two curves:

- The blue area corresponds to the profile of the variable selected over the cluster selected.
- The red area corresponds to the profile of the variable selected over the entire dataset.

The figure below presents the Cross Statistics obtained in this scenario for cluster 3.
In the figure above, the table allows you to identify cluster 3 as the cluster containing the highest density of observations belonging to the target category of the target variable: 75% of customers contained in this cluster belong to target category 1 of the target variable Class.

The cross statistics plot allows you to view and compare the profiles of the variable capital-gain over the entire dataset and over cluster 3. These profiles are repeated in the table below.

<table>
<thead>
<tr>
<th>Categories of the variable “capital-gain”</th>
<th>Profile over the dataset</th>
<th>Profile over cluster 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>KxMissing</td>
<td>1%</td>
<td>7%</td>
</tr>
<tr>
<td>[0]</td>
<td>92%</td>
<td>0%</td>
</tr>
<tr>
<td>[1409; 2964]</td>
<td>1%</td>
<td>0%</td>
</tr>
<tr>
<td>] 2964 ; 15024]</td>
<td>6%</td>
<td>88%</td>
</tr>
<tr>
<td>] 15024; 41310]</td>
<td>0%</td>
<td>5%</td>
</tr>
</tbody>
</table>

The data distribution of the category ] 2964 ; 15024] makes it clear that the majority of customers contained in cluster 3 realize significant annual capital gains relative to the entire set of customers contained in the dataset. In addition, the data distribution over the category [0] indicates that the majority of the customers contained in the dataset, or 92%, do not realize any annual capital gains, while none of the customers contained in cluster 3 fail to realize some annual capital gain.

Checking the Fix Variable box would allow you to compare the profiles of the variable capital-gain for all the segments.

### 6.5.3.6.4 Displaying SQL Expressions

The Cross Statistics screen also allows you to visualize the SQL Expression used to define each cluster.

**Note**

SQL expressions are only available if you have selected the Calculate SQL Expressions option in the modeling parameters advanced screen before generating your model.

To Display SQL Expression for a Cluster

1. Select the cluster in the summary table.
   
   The plot for the selected cluster is displayed.

2. Click View Type and select the SQL button.
   
   The SQL expression replaces the cross statistics plot in the lower part of the screen.

3. Click + or - to fold or unfold the tree representing the SQL expression.

4. Click View Type and select Profiles to go back to the Cross Statistics plot.
6.5.3.6.5  Understanding SQL Expressions

The SQL Expressions screen can be broken down into two parts:

- In the upper part, a table presents each cluster in a summarized fashion. It allows you to select the cluster for which you want to display a SQL expression.
- In the lower part, a tree presents the SQL expression corresponding to the selected cluster.

The following schema presents the SQL expression for Cluster 6.

```
AND
  - age in [17 : 27]
  - marital-status in (Divorced, Married-spouse-absent, Never-married, Separated, Widowed)
  - occupation not in (Exec-managerial, Prof-specialty)

NOT
  - capital-gain in (KxMissing or [2964:41310])
  - capital-loss in ([1408; 1504], [1721; 4356])
```

The SQL expression can be broken down as follows:

- The first part (1) defines a cluster of observations where the variables equal the values displayed.
- The second part (2) defines clusters of observations that are excluded for the cluster found in part 1. The percentages displayed indicate the proportion of each cluster excluded with respect to the cluster found in part 1.

In our example, the first excluded cluster corresponds to observations where the value of capital-gain ranges between 2964 excluded and 41310 ([2964; 41310]), or is missing (KxMissing). It represents 0.21% of the observations found in part 1.

Note

The clusters are created by applying the SQL expressions in a specific order defined by the engine. If you apply the SQL rules randomly, you may not obtain exactly the same result.

6.5.3.6.6  Difference Between Standard Cross Statistics and SQL Expressions

When you ask for SQL expressions, the final segmentation is different from the one without. The goal of SQL is to have easy-to-understand and easy-to-apply segments. SQL expressions are built to describe as much as possible the basic segments (that is the ones you get when you do not ask for SQL). The SQL can be used both to have a better definition/understanding of the clusters and to deploy them on the full database or on new data (which is not usually trivial with other techniques).

The best way to understand the difference between centroid-based clusters and rule-based clusters is to use graphs.
This diagram represents a set of observations from a dataset.

To create clusters, Modeler - Segmentation/Clustering engine uses the centroid approach. Centroids are the results of a clustering algorithm, meaning they are the barycenter of the points closest to them. When applying Modeler - Segmentation/Clustering on this dataset, the observations are grouped depending on their distance with each centroid.

This graph represents the previous dataset observations grouped into four clusters.

This diagram is known as a Voronoi diagram.

To create the SQL expressions that define the clusters, Modeler - Segmentation/Clustering engine uses what is called Minimum Description Length (MDL). It means that after creating the initial clusters from the centroid approach then they are reshaped, cut to fit into the smallest possible expression thus trying to find the best compromise between length of the expression and the lost of information.

This graph represents the SQL expressions of the clusters (in red) compared with the centroids.

You can see on this graph that:

- some observations that were in a cluster when using the centroid approach end up in another when using the SQL expressions.
- some observations can not be described by the SQL expressions and are left outside the cluster. They are called the unassigned observations.
- some observations are described by two different SQL expressions, thus appearing in two clusters. This is called the overlap.
This graph presents the final result obtained with SQL expressions.

An observation cannot appear in two different clusters, so when there is overlap between clusters, the observation concerned by the overlap is kept in the first cluster created. The second cluster that was also containing the observation is redefined to exclude it. In this schema, the numbers correspond to the order of creation of the clusters.

You can see that the observations that were in two clusters are kept in only one. The choice of the cluster in which the overlapping observations are kept depends on the order in which the SQL rules are applied. In this case, the rule defining cluster 2 has been applied before the rules defining the clusters 1 and 3.

How to decide which segmentation is better

As a side effect of the supervision, Segmentation/Clustering provides you with a KI and KR. It can be used to compare the two segmentations (especially because the number of segments is the same). If KI does not change significantly, then the one with SQL may be preferred because it is easier to understand. If there is a fall of KI, you may want to stick with the basic segmentation.

KI may not be the thing you want to optimize for segmentation. The target profile of each segment is available in the GUI. Out of the four clusters, maybe one or two are of real interest. In that case you have to focus on these interesting segments and see how they evolve with SQL generation.

6.5.3.7 Statistical Reports

Statistical Reports provide you with a set of tables that allow you a more detailed debriefing of your model. These reports are grouped in different levels of debriefing:

- The Descriptive Statistics, which provides the statistics on the variables, their categories and the datasets, as well as the variables cross-statistics with the target.
**Note**

If your dataset contains date or datetime variables, automatically generated variables will appear in the statistical reports.

- The Model Performance, in which you will find the model performance indicators, the variables contributions and the score detailed statistics.
- The Control for Deviations, which allows you to check the deviations for each variable and each variable category between the validation and test datasets.
- The Expert Debriefing, in which you will find more specialized performance indicators, as well as the variables encoding, the excluded variables during model generation and the reason for exclusion, and so on.

**Related Information**

Variable Storage Formats [page 30]

### 6.5.3.71 Statistical Report Options

A toolbar is provided allowing you to modify how the current report is displayed, to copy the report, to print it, to save it or to export it to Excel.

**Display Options**

<table>
<thead>
<tr>
<th>Menu</th>
<th>Option</th>
</tr>
</thead>
<tbody>
<tr>
<td>View</td>
<td><img src="image" alt="Display the current report view in the graphical table that can be sorted by column." /></td>
</tr>
<tr>
<td></td>
<td><img src="image" alt="Display the current report view as an HTML table." /></td>
</tr>
<tr>
<td></td>
<td><img src="image" alt="Some reports can be displayed as a bar chart. This bar chart can be sorted by ascending or descending values, or by ascending or descending alphabetical order. You can also select which data should be displayed." /></td>
</tr>
<tr>
<td></td>
<td><img src="image" alt="Display certain reports as a pie chart." /></td>
</tr>
<tr>
<td></td>
<td><img src="image" alt="Display certain reports as a line chart." /></td>
</tr>
<tr>
<td>Sort</td>
<td><img src="image" alt="When the current report is displayed as a bar chart, this option allows you to change the orientation of the bars (from horizontal to vertical and vice versa)." /></td>
</tr>
</tbody>
</table>
### Menu

<table>
<thead>
<tr>
<th>Option</th>
</tr>
</thead>
<tbody>
<tr>
<td>Display the current report with no sorting.</td>
</tr>
<tr>
<td>Sort the current report by ascending values.</td>
</tr>
<tr>
<td>Sort the current report by descending values.</td>
</tr>
<tr>
<td>Sort the current report by ascending names.</td>
</tr>
<tr>
<td>Sort the current report by descending names.</td>
</tr>
</tbody>
</table>

#### Series

Select which columns to display for current report.

### Usage Options

<table>
<thead>
<tr>
<th>Option</th>
</tr>
</thead>
<tbody>
<tr>
<td>Copy the data from the current view of the displayed report. The data can then be pasted in a text editor, a spreadsheet, a word processing software.</td>
</tr>
<tr>
<td>Print the current view of the selected report depending on the chosen display mode (HTML table, graph, ...).</td>
</tr>
<tr>
<td>Save under different formats (text, html, pdf, rtf) the data from the current view of the selected report.</td>
</tr>
<tr>
<td>Save under different formats (text, html, pdf, rtf) the data from all the views of the selected report.</td>
</tr>
<tr>
<td>Export to Excel.</td>
</tr>
<tr>
<td>Save all reports.</td>
</tr>
<tr>
<td>Save the customized style sheet.</td>
</tr>
</tbody>
</table>

### 6.5.4 Step 4 - Using the Model

Once generated, a clustering model may be saved for later use. A clustering model may be applied to additional datasets. The model thus allows you to assign observations to clusters. This part presents the option of
applying the model to a new dataset for the Segmentation/Clustering feature. The other options for deployment of the clustering models are similar to those proposed for models generated using the Regression/Classification feature. For more information about these options, see the topics on saving and opening a model.

Applying the Model to a New Dataset

The currently open model may be applied to additional datasets. The model allows you to determine to which cluster the observations described in these datasets belong.

Constraints of Model Use

In order to apply a model to a dataset, the format of the application dataset must be identical to that of the training dataset used to generate the model. The same target variable, in particular, must be included in both datasets, even if values for the target variable are not contained in the application dataset.

Related Information

Saving the Model [page 157]
Opening a Model [page 159]

6.5.4.1 Using the Option Direct Apply in the Database

This optimized scoring mode can be used if all the following conditions are met:
- the apply-in dataset (table, view, select statement, data manipulation) and the results dataset are tables coming from the same database,
- the model has been computed while at least one physical key variable was defined in the application,
- no error has occurred,
- the in-database apply mode is not deactivated,
- granted access to read and write (create table).

To use the in-database apply mode, check the option Use the Direct Apply in the Database and automatically the option Add Score Deviation is selected as well.
6.5.4.2 Advanced Apply Settings

6.5.4.2.1 General Outputs

Copy the Weight Variable

This option allows you to add to the output file the weight variable if it had been set during the variable selection of the model.

Copy Data Set Id

This option allows you to add to the output file the name of the sub-data set the record comes from (Estimation, Validation or Test).

Note

This option cannot be used with the in-database apply feature.

Copy the Variables

This option allows you to add to the output file one or more variables from the data set.

- To Add All the Variables
  - Check the All option.
- To Select only Specific Variables
  1. Check the Individual option.
  2. Click the >> button to display the variable selection table.
  3. In the Available list select the variables you want to add (use the Ctrl key to select more than one variable).
  4. Click the > button to add the selected variables to the Selected list.

User Defined Constant Outputs

This option allows you to add to the output file constants such as the apply date, the data set name, or any other information useful for using the output file.
A user defined constant is made of the following information:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Value / Warnings</th>
</tr>
</thead>
</table>
| Visibility  | indicates if the constant will appear in the output or not | *checked*: the constant appears in the output  
*unchecked*: the constant does not appear in the output |
| Name        | the name of the user defined constant     | ● The name cannot be the same as the name of an existing variable of the reference data set.  
● If the name is the same as an already existing user defined constant, the new constant will replace the previous one |
| Storage     | the constant type (number, string, integer, date, ...) | *number*  
*string*  
*integer*  
*date*  
*datetime* |
| Value       | the value of the constant                 | date format: YYYY-MM-DD  
datetime format: YYYY-MM-DD HH:MM:SS |
| Key         | indicates if the constant is a key variable or identifier for the record. You can declare multiple keys. They will be built according to the indicated order (1-2-3-...). | 0: the variable is not an identifier  
1: primary identifier  
2: secondary identifier  
... |

- To Define a Constant
  1. Click the *Add* button. A pop-up window opens allowing you to set the constant parameters.
  2. In the field *Output Name*, enter the constant name.
  3. In the list *Output Storage*, select the constant type.
  4. In the field *Output Value*, enter the constant value.
  5. Click the *OK* button to create the constant. The new constant appears in the list. You can choose whether to generate the defined constants or not by checking the *Visibility* box.
6.5.4.2.2 Outputs by Cluster Identifier, Cluster Rank, and Miscellaneous Outputs

6.5.4.2.2.1 For Clustering Models

Outputs by Cluster Ranks

Distance to Clusters

Use this option to add to the output file the distance of each observation from the clusters. The distances are generated in the columns named \texttt{kc\_dist\_cluster\_<TargetVariable>\_<ClusterId>}. For example, if the target variable is \textit{Age}, the distance from cluster 1 will appear in the column \texttt{kc\_dist\_cluster\_Age\_1}.

- To Add the Distances from All Clusters
  - Select \textit{All}.
- To Select Distances from Specific Clusters
  1. Select \textit{Individual}.
  2. Click the >> button to display the cluster selection table.
  3. Select the clusters for which you want to add the distance.

\textbf{i Note}

When SQL mode is activated, the notion of nearest cluster does not exist. If a case belongs to a cluster, distance is set to 0. If a case does not belong to a cluster, distance is set to 1.

Probability for Clusters

Use this option to add to the output file the probability of each observation to belong to the various clusters. The probabilities are generated in the columns \texttt{kc\_proba\_cluster\_<TargetVariable>\_<ClusterId>}. For example, if the target variable is \textit{Age}, the probability that the observation belongs to cluster 1 will be displayed in the column \texttt{kc\_proba\_cluster\_Age\_1}.

- To Add the Probabilities for All Clusters
  - Select \textit{All}.
- To Select the Probabilities for Specific Clusters
  1. Select \textit{Individual}.
  2. Click the >> button to display the cluster selection table.
  3. Select the clusters for which you want to add the probabilities.

\textbf{i Note}

When SQL mode is activated, the notion of nearest cluster does not exist. If a case belongs to a cluster, probability is set to 1. If a case does not belong to a cluster, probability is set to 0.
Outputs by Cluster Identifier

Top Ranking Cluster Indices
Use this option to add to the output file the number of the clusters whose centroids are the closest to the current observation. The closest cluster is the one the observation belongs to, its number is displayed in the column \texttt{kc_<Target variable>}. The next closest cluster is displayed in the column \texttt{kc_<Target Variable>_2}, and so on until the furthest cluster. You can choose to add all the clusters or only the closest.

- To Add All the Clusters
  - Select \texttt{All}.
- To Add Only the Closest Clusters
  1. Select \texttt{Top}.
  2. In the text field enter the number of clusters you want to add; for example, the two, three or four closest clusters.

Top Ranking Cluster Names
Use this option to add to the output file the names of the clusters whose centroids are the closest to the current observation. The closest cluster is the one the observation belongs to, its name is displayed in the column \texttt{kc_name_<Target variable>}. The next closest cluster is displayed in the column \texttt{kc_name_<Target Variable>_2}, and so on until the furthest cluster. You can choose to add all the clusters or only the closest one.

- To Add All the Clusters
  - Select \texttt{All}.
- To Add Only the Closest Clusters
  1. Select \texttt{Top}.
  2. In the text field enter the number of clusters you want to add; for example, the two, three or four closest clusters.

\begin{itemize}
  \item Note
  
  The name of a cluster is its number by default. You can modify it in the column \texttt{User Name} of the panel \textit{Clusters Profiles} accessible through the main menu.
\end{itemize}

Top Ranking Distances
Use this option to add to the output file the distances of each observation from the clusters centroids. The distance from the closest centroid is displayed int the column \texttt{kc_best_dist_<TargetVariable>}, the distance from the second closest centroid is displayed in the column \texttt{kc_best_dist_<TargetVariable>_2}, and so on until the furthest centroid. You can add the distances from all centroids or only the shortest.

- To Add All the Distances
  - Select \texttt{All}.
- To Add Only the Shortest Distances
  1. Select \texttt{Top}.
  2. In the text field enter the number of distances you want to add; for example, the two, three or four shortest distances.
**Note**

When SQL mode is activated, the notion of nearest cluster does not exist. If a case belongs to a cluster, distance is set to 0. If a case does not belong to a cluster, distance is set to 1.

**Probabilities**

Use this option to add to the output file the probabilities that the observation belongs to each cluster. The probability for the observation to belong to the closest cluster is displayed in the column `kc_best_proba_<TargetVariable>`, this probability is usually the highest. The probability for the observation to belong to the second closest cluster is displayed in the column `kc_best_proba_<TargetVariable>_2`, and so on until the furthest cluster. You can add all the probabilities or only the ones corresponding to the closest clusters.

- To Add All Probabilities
  - Select *All*.
- To Add Only the Probabilities for the Closest Clusters
  1. Select *Top*.
  2. In the text field enter the number of probabilities you want to add; for example, the two, three or four best probabilities.

**Note**

When SQL mode is activated, the notion of nearest cluster does not exist. If a case belongs to a cluster, probability is set to 1. If a case does not belong to a cluster, probability is set to 0.

**Miscellaneous Outputs**

**Disjunctive Coding**

Use this option to add to the output file the disjunctive coding of the clusters. A column is generated for each cluster and contains either 0 or 1 depending whether the observation belongs to the cluster or not. The columns created are named `kc_disj_<TargetVariable>_<ClusterId>`. For example if the target variable is Age and the model has five clusters, the five following columns will be generated `kc_disj_age_1, kc_disj_age_2, kc_disj_age_3, kc_disj_age_4, kc_disj_age_5`.

**Target Mean / Target Key Probability**

Use this option to add to the output file:

- for continuous targets:
  - the mean of the target for the cluster containing the observation (displayed in the column `kc_<TargetVariable>_Mean`).
  - the difference with the actual target value if the latter is known for the current observation (displayed in the column `kc_<TargetVariable>_Error`).

  Note that when the actual target value is not available, it is set to 0. Thus, the difference between the actual target value and the computed target mean equals to the same value, meaning that both `kc_<TargetVariable>_Mean` and `kc_<TargetVariable>_Error` columns display the same value.
for nominal targets:

- the proportion of the least frequent category of the target variable (key category) in the cluster containing the current observation (displayed in the column \( kc_{<\text{TargetVariable}>\_Mean} \)).

### 6.5.4.3 Types of Results Available

The application of a model to a dataset allows you to obtain three types of results:

- The cluster index for each observation.
- The disjunctive encoding (or dummy coding) of the cluster indexes, which means that, for each cluster, a boolean variable is created indicating whether the current observation belongs to that cluster or not. For a given observation, the value “1” is assigned to the variable corresponding to the cluster containing the observation and the value “0” is assigned to the variables corresponding to the other clusters. The variable names are built according to the following pattern: \( kc_{<\text{TargetName}>\_<\text{ClusterIndex}>} \). Consider, as an example, that you have generated a five-clusters model. When applying this model, the application creates five variables corresponding to the five generated clusters. For an observation belonging to cluster 3, the result appears as shown below.

<table>
<thead>
<tr>
<th>KxIndex</th>
<th>class</th>
<th>kc_class</th>
<th>kc_class_1</th>
<th>kc_class_2</th>
<th>kc_class_3</th>
<th>kc_class_4</th>
<th>kc_class_5</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

- The target mean for each cluster, that is, the percentage of observations belonging to the target category of the target variable contained in each cluster.

Depending upon the level of information desired, you can choose to generate:

- Only the cluster index to which each observation belongs (Predicted Value Only option).
- The cluster index and the disjunctive encoding of the cluster indexes (Cluster Id Disjunctive Coding option).
  You can also decide to include in the results file all input variables of the application dataset (Cluster Id Disjunctive Coding (+copy dataset) option).
- The cluster index and the target mean for each cluster (Cluster Id Target Mean option).

For this scenario, you will apply the model to the file Census01.csv that you used previously to generate the model.

In the procedure to apply a model to a new dataset:

- Select the format Text files.
- In the Generate field, select the option Cluster Id Target Mean.
- Select the folder of your choice in which to save the results file (Results generated by the model).

### Related Information

Applying the Model to a New Dataset [page 134]
6.5.4.4 Analyzing the Results of the Application

For this scenario, open the results text file in Microsoft Excel, generated when you applied the model to the Census01.csv file.

To open the model application results file:

1. Depending upon the format of the results file generated, use Microsoft Excel or another application to open the file.

The figure below presents the headings and columns of the results file obtained for this scenario.

2. You can now analyze the results obtained and use these results of your analysis to make the right decisions.

6.5.4.4.1 Description of the Results File

Depending upon which options you selected, the results file will contain some or all of the following information, in the same order as seen below:

- The key variable defined during data description at the model parameter settings step. If your dataset did not contain a key variable, the key variable \(KxIndex\) would have been generated automatically by the application.
- Possibly, the target variable given as known values if the latter appeared in the application dataset, as is the case in this scenario.
- The variable \(kc\_clusterId\), which indicates the number of the cluster to which each observation belongs.
- The variable \(kc\_TargetMeanClusterId\), which indicates the proportion of observations belonging to the target category of the target variable that are contained in each cluster.
- The variables corresponding to each cluster, and an indication of the encoding disjunction of the cluster numbers. The names of these variables correspond to cluster numbers, prefixed by \(kc\_cluster\_\)_, for example \(kc\_cluster\_1\) for cluster 1.
7 Glossary

7.1 analytical data management

The analytical data management as defined by the application is made of three elements:
- the data manipulation functions offered by the application, such as filters, join attributes, new attributes computation, aggregates, performance indicators definition,
- the analytical dataset methodology,
- the meta data management, which allows storing, sharing and easily re-using the data descriptions.

7.2 analytical dataset

Tabular representation of data made of lines and columns. Each line represents an “observation”. Roles can be assigned to columns such as “Input”, “skip”, “target” or “weight”.

7.3 analytical record

An analytical record is a logical view of all attributes corresponding to an entity. An analytical record may be decomposed into domains that group attributes related to each other: for example, in CRM, an analytical record can have a demographic domain and a behavioral domain.

7.4 antecedent

X is called the antecedent of the rule. The antecedent can be composed of an item or an itemset, for example X can be the set \{A,B,C\}.
7.5 application dataset

An application dataset is a dataset to which you apply a model. This dataset contains an unknown target variable for which you want to know the value.

7.6 association rule

An association rule is an implication relation of the form $X \Rightarrow Y$. The rule means: if the attribute $X$ is present in a session, then the attribute $Y$ is present too. Two measures allow qualifying the quality of the rule: the Support and the Confidence.

7.7 attribute

In computing, an attribute is a specification that defines a property of an object, element, or file.

7.8 AUC

The AUC statistic is a rank-based measure of model performance or predictive power calculated as the area under the Receiver Operating Characteristic curve. For a simple scoring model with a binary target, this represents the observed probability of a signal (responder) observation having a higher score than a non-signal (non-responder) observation. For individual variables, ordering based on score is replaced by ordering based on the response probability for the variable’s categories (for example cluster ID or age range response rates).

7.9 authenticated server

Users will be able to communicate to the application authenticated server only when providing correct password. Authenticated server delegates the authentication to Custom built services or Operating System services through PAM (Pluggable Authentication Modules).
7.10 **autoselection**

It is an automated attribute selection.

7.11 **bin**

A bin is a range of values defined by its bounds (upper bound and lower bound). Bins result from a data manipulation activity known as binning. Synonym: range.

7.12 **bipartite graph display / non-bipartite graph display**

The bipartite graph display shows two distinct populations of nodes (or node sets) with the links between the two node sets. For example, the first node set could represent clients and the second, products. From this global view, a non-bipartite graph display can be derived to focus on the links between the nodes of a given node set.

7.13 **bubble chart**

A bubble chart is a specific graphical representation in Modeler - Segmentation/Clustering which displays clusters as bubbles. The coordinates of a given bubble are the cluster centroid values according to two selectable continuous variables. The size of the bubble is plotted according to the frequency of the corresponding cluster.

7.14 **calendar table**

A calendar table is used to ease the development of solutions around any business model which involves dates. A common practice is to have a calendar table pre-populated with some or all of the needed information, enabling to accomplish most date related complex tasks with simple database queries.
7.15 category

A category is one of the possible values of a discrete variable. A discrete variable is a nominal or ordinal variable. It is the basic element used to code the variable as well as to gather descriptive statistics.

7.16 category significance

The category significance measures the impact a category has on the target.

7.17 centroid

Imaginary point inside a polygon whose coordinates are generally those of the polygon center.

7.18 chunk (by chunk)

Number of lines of a table that are processed as package.

7.19 classification rate

The ratio between the number of correctly classified records and the total number of records.

7.20 confidence

The Confidence of a rule is a measure that indicates the percentage of sessions verifying the consequent among those verifying the antecedent. For instance the number of sessions containing the item D, among the ones containing the itemset \{A,B,C\}. 
7.21  confusion matrix

The confusion matrix allows visualizing the target values predicted by the model compared with the real values and setting the score above which the observations will be considered as positive, that is the observations for which the target value is the one wanted.

7.22  consequent

Y is called the consequent of the rule. The consequent is composed of only one item, for example Y can be the item (D).

7.23  continuous variable

Continuous variables are variables whose values are numerical, continuous and sortable. Arithmetic operations may be performed on these values, such as determination of their sum or their mean.

7.24  contribution

Relative importance of each variable in the built model.

7.25  correlation

Any measure that quantifies the fact that two variables share the same information. This can be measured by looking at the relative variation of the two variables for different entities. Classical statistics defines linear correlation to compute such a metrics on continuous variables. The application computes correlations between variables of different types by looking at the correlation of the codes of both variables in presence of a target.
7.26 cross statistics

A method of estimating the accuracy of a classification or regression model. The dataset is divided into several parts, with each part in turn used to test a model fitted to the remaining parts.

7.27 customized cutting strategy

The customized cutting strategy allows you to define your own data sub-sets. To use this strategy, you must have prepared (before opening the application features) three sub-sets: the estimation, validation and test sub-sets.

7.28 customized profit

Customized profit allows you to define your own profit values, that is, to associate both a cost and a benefit to each value of the target variable.

7.29 cutting strategy

A cutting strategy is a technique that allows decomposition of a training dataset into two or three distinct sub-sets:

- An estimation sub-set,
- A validation sub-set,
- A test sub-set.

This cutting allows for cross-validation of the models generated.

7.30 data aggregation

The process of consolidating data values into a smaller number of values. For example, sales data could be collected on a daily basis and then be totaled to the week level.
7.31 dataset

A collection of data, usually presented in tabular form, where each column represents a particular variable and where each row is an assignment of values.

7.32 data source

A data source includes both the source of data itself, such as relational database, a flat-file database, or even a text file, and the connection information necessary for accessing the data.

7.33 database

A database is a structured collection of records or data that is stored in a computer system.

7.34 descriptive model

A model which allows describing datasets.

7.35 detected profit

Detected profit is the profit type shown as the default. It allows examination of the percentage of observations belonging to the target category of the target variable, that is, the least frequent category, as a function of the proportion of observations selected from the entire dataset.

7.36 determination coefficient (R2)

Ratio between the variability (sum of squares) of the prediction and the variability (sum of squares) of the data.
7.37 deviation

Deviation is a measure of difference for interval and ratio variables between the observed value and the mean.

7.38 domain

See Analytical Record. The behavioral domain is usually obtained through aggregates per entity on transactional tables.

7.39 encoding

Encoding is the process of putting a sequence of characters (letters, numbers, punctuation, and certain symbols) into a specialized format for efficient transmission or storage.

7.40 engine

The UI and view independent portion of an application, concerned with data manipulation and other fundamental operations independently of how these are eventually represented to the user.

7.41 entity

An entity is the object of interest of any analytical task: it can be a customer, a product, a store, and is usually identified with a single identifier that can be used throughout the data repositories. Entities are usually associated with a state model describing the life cycle of such an analytical object of interest.

i Note

This is a technical constraint: entities MUST be uniquely identified.
7.42 error bar

See prediction range

7.43 error mean

Mean of the difference between predictions and actual values

7.44 error standard deviation

Dispersion of errors around the actual result

7.45 event dataset

An event dataset should consist of at least:
• An event date such as birthdate or beginning of trial in YYYY/MM/DD format
• A reference id (i.e. customer id), that will be used to join the Events or transactions data with the reference or static customer table previously defined.

7.46 excluded variable

Actual target

7.47 explanatory variable

An explanatory variable is a variable that describes your data and which serves to explain a target variable.
7.48 Expression Editor

Panel allowing to create fields as complex expressions in the Analytical Data set Editor.

7.49 extra-predictable variable

Variable whose values are known for the period that is to be predicted.

7.50 false positive

Incorrect assignments to the signal class

7.51 fluctuation

Evolution of the signal that is not stable neither cyclic (Modeler – Time Series).

7.52 GINI index

The GINI statistic is a measure of predictive power based on the Lorenz curve. It is proportionate to the area between the random line and the Model curve.

7.53 horizon-wide MAPE

This quality indicator for the forecasting model is the mean of MAPE values observed over all the training horizon. A value of zero indicates a perfect model while values above 1 indicate bad quality models. A value of 0.09 means that the model takes into account 91% of the signal or, in other words, the forecasting error (model residues) is relatively of 9%.
7.54 **In-database apply**

The in-database apply is used to apply a model into a database:

- The proper SQL code is generated for the model,
- The resulting code is then executed as a single SQL request in the database

This avoids extracting the data from the database, and speeds up the writing process of the model outputs.

7.55 **item**

A component of an association rule.

7.56 **itemset**

A group (or a set) of items, is called an itemset

7.57 **iteration**

An "iteration" is a single loop through a cycle, such as the design-prototype-test cycle.

7.58 **KL (Kullback-Leibler)**

The Kullback-Leibler divergence is used to measure the difference between the cluster profile and the population profile of the variables.

7.59 **KPI (Key Performance Indicator)**

KPIs, or key performance indicators help organizations achieve organizational goals through the definition and measurement of progress. The key indicators are agreed upon by an organization and are indicators which can
be measured that will reflect success factors. The KPIs selected must reflect the organization's goals, they must be key to its success, and they must be measurable.

### 7.60 K-S test

K-S is the Kolmogorov-Smirnov statistic applied here as a measure of deviation from uniform response rates across categories of a variable. Kolmogorov-Smirnov is a non-parametric, exact goodness-of-fit statistic based on the maximum deviation between the cumulative and empirical distribution functions.

### 7.61 Lift

The Lift of a rule is a measure that indicates the chances of finding the consequent by using the antecedent compared with the chances of randomly finding the consequent. A value greater than 1 indicates that using the antecedent increases your chances to find the consequent.

### 7.62 Lift profit

Lift profit allows examination of the difference between a perfect model and a random model and between the model generated by the application and a random model. It represents the ratio between a model and the random model, that is the performance of a model compared to a model that would only allow you to select observations at random from your database.

You can thus visualize how much better your model is compared with the random model.

### 7.63 lower bound

The term lower bound is defined as an element of $P$ which is lesser than or equal to every element of $S$. 
7.64 **maximum error (LInf)**

Maximum absolute difference between predicted and actual values (upper bound) (Chebyshev distance)

7.65 **mean**

The arithmetic average value of a collection of numeric data.

7.66 **mean absolute error (L1)**

Mean of the absolute values of the differences between predictions and actual results, (City block distance or Manhattan distance.)

7.67 **mean absolute percentage error (MAPE)**

The MAPE value is the average of the sum of the absolute values of the percentage errors. It measures the accuracy of the model's forecasts and indicates how much the forecasts differ from the real signal value.

7.68 **mean square error (L2)**

Square root of the mean of the quadratic errors (Euclidian distance or root mean squared error - RMSE)

7.69 **metadata**

The information about the data itself.
7.70 **meta-operator**

Operators that are used upon other operators.

7.71 **missing value**

Data values can be missing because they were not measured, not answered, were unknown or were lost.

7.72 **monotonicity**

The direction of variation of a monotonic function does not change.

7.73 **multiple-instance installation**

An installation mode that consists in running several instances on one server in order to divide up the load.

7.74 **nominal variable**

Nominal variables are variables whose values are discrete, that is, belong to categories, and are not sortable.

- Nominal variables may be:
  - Numerical, meaning that its values are numbers.
  - Textual, meaning that its values are character strings.

Important: binary variables are considered nominal variables.

7.75 **normalize**

To transform numerical data and make them fit into a defined interval.
7.76 numeric filter

A digital filter is a system that performs mathematical operations on a sampling, discrete-time signal to reduce or enhance certain aspects of that signal.

7.77 ordinal variable

Ordinal variables are variables with discrete values, that is, they belong to categories, and they are sortable. Ordinal variables may be:
- Numerical, meaning that its values are numbers. They are therefore ordered according to the natural number system (0, 1, 2, and so on).
- Textual, meaning that its values are character strings. They are therefore ordered according to alphabetic conventions.

7.78 outlier

A data value that does not come from the typical population of data; in other words, extreme values. In a normal distribution, outliers are typically at least 3 standard deviations from the mean.

7.79 performance indicator (PI)

Performance indicators help organizations achieve organizational goals through the definition and measurement of progress. The purpose of defining PIs is to have a common definition of a metric across multiple projects. A metric like “customer value” could easily be defined in several different ways, leading to confusing or contradictory results from one analysis to the next. Shared PIs ensure consistency across analysts and projects over time. The key indicators are agreed upon by an organization and are indicators which can be measured and will reflect success factors. The PIs selected must reflect the organization’s goals, they must be the key to its success, and they must be measurable.
7.80 periodic cutting strategy

The periodic cutting strategy is implemented by following this distribution cycle:

- Three lines of the initial dataset are distributed to the estimation sub-set.
- One line is distributed to the validation sub-set.
- One line is distributed to the test sub-set.
- Distribution begins again at step 1.

7.81 pivot

A pivot is a data summarization tool found in data visualization programs. Among other functions, they can automatically sort, count, and total the data stored in one table or spreadsheet and create a second table displaying the summarized data. Pivot tables are also useful for quickly creating cross tabs.

7.82 polynomial

A polynomial may be of degree 1, 2, 3 or greater. By defining the polynomial degree, you are defining the degree of complexity of the model.

7.83 population

A population is a list of entity identifiers. A population may be defined as list of values. This list can be extracted from a column table: it is then said to be defined in extension, or through a filtering expression from another population: it is then said to be defined in intension.

7.84 prediction range

The extreme values for prediction ranges are \( \text{TargetMean} - \sqrt{\text{TargetVariance}}; \text{TargetMean} + \sqrt{\text{TargetVariance}} \).
7.85  predictive model

A model which allows predicting phenomena.

7.86  profit type

A profit type allows calculation of the profit that may be realized using the model. In general, a benefit is associated with the positive (or expected) values of the target variable and a cost is associated with the negative (or unexpected) values.

7.87  quality indicator: the predictive power (KI)

The predictive power is the quality indicator of the models generated using the application. This indicator corresponds to the proportion of information contained in the target variable that the explanatory variables are able to explain.

7.88  random cutting strategy

The random cutting strategy distributes the data of the initial dataset in a random manner between the three sub-sets, estimation, validation and test.

7.89  record

The fundamental data structure used for performing data analysis. Also called a table row or example. A typical record would be the structure that contains all relevant information pertinent to one particular customer or account.
7.90 robustness

The degree of robustness corresponds to the predictive power of the model applied to an application dataset.

7.91 robustness indicator: the prediction confidence (KR)

The prediction confidence is the robustness indicator of the models generated using the application. It indicates the capacity of the model to achieve the same performance when it is applied to a new dataset exhibiting the same characteristics as the training dataset.

7.92 ROC

The ROC (Receiver Operating Characteristic) graph is derived from signal detection theory. It portrays how well a model discriminates in terms of the tradeoff between sensitivity and specificity, or, in effect, between correct and mistaken detection as the detection threshold is varied.

7.93 role

In data modeling, variables (see page 288) may have three roles. They may be:

- Target variables,
- Explanatory variables,
- Weight variables.

7.94 root

Terminological morpheme that is used alone as word - root word - or as basic element in a derived word.
7.95 **score**

The numeric evaluation mark in view of a given problem.

7.96 **scorecard**

This screen provides you with the coefficients associated to each category for all variables in the model (only in case of a regression model (Classification/Regression)).

7.97 **seasonal**

Variations due to calendar events.

7.98 **sensitivity**

Sensitivity, which appears on the Y axis, is the proportion of correctly identified signals (true positives) found (out of all true positives on the validation dataset).

7.99 **sequential cutting strategy**

The sequential strategy cuts the initial dataset into three blocks, corresponding to the usual cutting proportions:

- The lines corresponding to the first 3/5 of the initial dataset are distributed as a block to the estimation dataset.
- The lines corresponding to the next 1/5 of the initial dataset are distributed as a block to the validation dataset.

The lines corresponding to the final 1/5 of the initial dataset are distributed as a block to the test dataset.
7.100 session

A session is identified by a unique key and is composed by one or more transactions.

7.101 simulation

Application of a model to only one record.

7.102 smart variable contributions

The variable contribution in a model while taking into account the variable correlation.

7.103 social network analysis

Social network analysis is used to approach problems such as community identification, diffusion in graphs (product adoption, epidemiology), graph evolution or influence of an individual within a community (leader vs. follower).

7.104 standard deviation

The standard deviation is a measure of the dispersion of a collection of numbers.

7.105 standardized profit

Standardized profit allows examination of the contribution of the model generated by v features relative to a model of random type, that is, in comparison with a model that would only allow to select observations at random from your database.
This profit is used for the plots of variable details, which present the significance of each of the categories of a given variable with respect to the target variable.

### 7.106 statistical report

The Statistical Reports provide you with a set of tables that allows you a more detailed debriefing of your model.

### 7.107 storage

To describe the data, the application uses five types of storage formats:

- date
- datetime
- number
- integer
- string

### 7.108 sub-sampling

Sub-sampling means selecting a part of a whole: if an event cannot be processed as a whole, a limited number of measures have to be taken to represent this event.

### 7.109 support

The support of a rule is a measure that indicates the number of sessions that verify the rule. For instance the number of sessions, that contains the itemset \{A,B,C\} and the item D.

### 7.110 table of data

A table of data is a dataset presented in the form of a two-dimensional table.
7.111 target key

The target key is the expected value of the target.

7.112 target key

The target key is the expected value of the target.

7.113 target variable

A target variable is the variable that you seek to explain, or for which you want to predict the values in an application dataset. It corresponds to your domain-specific business issue.

7.114 temporal analytical dataset

A temporal analytical dataset is a special case of analytical dataset. It is the product of a time-stamped population by an analytical record, the result of this operation can be seen as a virtual table containing attributes values associated with identifiers in relation with the time stamp.

In other words: a temporal analytical dataset contains ‘photos’ (or snapshots), of a given list of entities taken at a given time, this time can be different for each entity, and an entity can be associated with several ‘photos’.

i  Note

Analytical datasets are used to train predictive/descriptive models and to apply these models.

7.115 time series

A time series is a sequence of data points, measured typically at successive times, spaced at (often uniform) time intervals.
7.116 timeout

A specified period of time that will be allowed to elapse before a specified event is to take place, unless another specified event occurs first.

7.117 time-stamped population

A time-stamped population is a list of pairs <identifiers; time stamps>: the semantic meaning of such a construct can be associated with snapshots of the entities and a given time; in general terms, a given entity may be represented at different time stamps in a single time stamped population.

7.118 training

Another term for estimating a model’s parameters based on the dataset at hand.

7.119 training dataset

A training dataset is a dataset used for generating a model. By analyzing the training dataset the application features will generate a model that allows explanation of the target variable, based on the explanatory variables.

7.120 transaction

A transaction is defined by the following:

- A unique key,
- The key of the related session,
- An attribute, called an item.
7.121 true negative
Correct assignments to the class of non-signals.

7.122 true positive
Correctly identified signal.

7.123 unassigned record
When creating clusters with SQL expressions, unassigned records are the observations that cannot be
described by the SQL expressions and are left outside the cluster.

7.124 upper bound
An upper bound of a subset $S$ of some partially ordered set $(P, \leq)$ is an element of $P$ which is greater than or
equal to every element of $S$.

7.125 variable
A variable corresponds to an attribute which describes the observations stored in your database.
In the application components, a variable is defined by: Type, Storage format, Role.

7.126 variable pool
The variable pool is a repository where the user stores the description of the frequently used variables. It is
located in the connector store (needs to be associated before, does nothing if not associated). This call stores
all the variable information usually edited by the user (standard description: storage, type value, etc, mapping
information and structure). The information stored in the pool is retrieved on the next guessed description. The user can also choose to save only the description of a specific variable.

7.127 variable type

There are four types of variables:
- continuous variables
- ordinal variables
- nominal variables
- textual variables

7.128 weight variable

A weight variable allows one to assign a relative weight to each of the observations it describes, and actively orient the training process. To declare a variable a weight variable results in creating a number of copies of each of the dataset observations, proportional to the value they possess for that variable.
Important Disclaimers and Legal Information

Coding Samples

Any software coding and/or code lines / strings ("Code") included in this documentation are only examples and are not intended to be used in a productive system environment. The Code is only intended to better explain and visualize the syntax and phrasing rules of certain coding. SAP does not warrant the correctness and completeness of the Code given herein, and SAP shall not be liable for errors or damages caused by the usage of the Code, unless damages were caused by SAP intentionally or by SAP’s gross negligence.

Accessibility

The information contained in the SAP documentation represents SAP’s current view of accessibility criteria as of the date of publication; it is in no way intended to be a binding guideline on how to ensure accessibility of software products. SAP in particular disclaims any liability in relation to this document. This disclaimer, however, does not apply in cases of willful misconduct or gross negligence of SAP. Furthermore, this document does not result in any direct or indirect contractual obligations of SAP.

Gender-Neutral Language

As far as possible, SAP documentation is gender neutral. Depending on the context, the reader is addressed directly with "you", or a gender-neutral noun (such as "sales person" or "working days") is used. If when referring to members of both sexes, however, the third-person singular cannot be avoided or a gender-neutral noun does not exist, SAP reserves the right to use the masculine form of the noun and pronoun. This is to ensure that the documentation remains comprehensible.

Internet Hyperlinks

The SAP documentation may contain hyperlinks to the Internet. These hyperlinks are intended to serve as a hint about where to find related information. SAP does not warrant the availability and correctness of this related information or the ability of this information to serve a particular purpose. SAP shall not be liable for any damages caused by the use of related information unless damages have been caused by SAP’s gross negligence or willful misconduct. All links are categorized for transparency (see: http://help.sap.com/disclaimer).