IBM SPSS Data Preparation 24

Note
Before using this information and the product it supports, read the information in "Notices" on page 33.

Product Information

This edition applies to version 24, release 0, modification 0 of IBM® SPSS® Statistics and to all subsequent releases and modifications until otherwise indicated in new editions.

Contents

Chapter 1. Introduction to Data Preparation

As computing systems increase in power, appetites for information grow proportionately, leading to more and more data collection—more cases, more variables, and more data entry errors. These errors are the bane of the predictive model forecasts that are the ultimate goal of data warehousing, so you need to keep the data "clean." However, the amount of data warehoused has grown so far beyond the ability to verify the cases manually that it is vital to implement automated processes for validating data.

The Data Preparation add-on module allows you to identify unusual cases and invalid cases, variables, and data values in your active dataset, and prepare data for modeling.

Usage of Data Preparation Procedures

Your usage of Data Preparation procedures depends on your particular needs. A typical route, after loading your data, is:

- **Metadata preparation.** Review the variables in your data file and determine their valid values, labels, and measurement levels. Identify combinations of variable values that are impossible but commonly miscoded. Define validation rules based on this information. This can be a time-consuming task, but it is well worth the effort if you need to validate data files with similar attributes on a regular basis.
- **Data validation.** Run basic checks and checks against defined validation rules to identify invalid cases, variables, and data values. When invalid data are found, investigate and correct the cause. This may require another step through metadata preparation.
- **Model preparation.** Use automated data preparation to obtain transformations of the original fields that will improve model building. Identify potential statistical outliers that can cause problems for many predictive models. Some outliers are the result of invalid variable values that have not been identified. This may require another step through metadata preparation.

Once your data file is "clean," you are ready to build models from other add-on modules.

© Copyright IBM Corporation 1989, 2016

Chapter 2. Validation Rules

A rule is used to determine whether a case is valid. There are two types of validation rules:

- **Single-variable rules.** Single-variable rules consist of a fixed set of checks that apply to a single variable, such as checks for out-of-range values. For single-variable rules, valid values can be expressed as a range of values or a list of acceptable values.
- **Cross-variable rules.** Cross-variable rules are user-defined rules that can be applied to a single variable or a combination of variables. Cross-variable rules are defined by a logical expression that flags invalid values.

Validation rules are saved to the data dictionary of your data file. This allows you to specify a rule once and then reuse it.

Load Predefined Validation Rules

You can quickly obtain a set of ready-to-use validation rules by loading predefined rules from an external data file included in the installation.

To Load Predefined Validation Rules

1. From the menus choose:

   Data > **Validation** > **Load Predefined Rules...**

Alternatively, you can use the Copy Data Properties Wizard to load rules from any data file.

Define Validation Rules

The Define Validation Rules dialog box allows you to create and view single-variable and cross-variable validation rules.

To Create and View Validation Rules

1. From the menus choose:

   Data > **Validation** > **Define Rules...**

   The dialog box is populated with single-variable and cross-variable validation rules read from the data dictionary. When there are no rules, a new placeholder rule that you can modify to suit your purposes is created automatically.
2. Select individual rules on the Single-Variable Rules and Cross-Variable Rules tabs to view and modify their properties.

Define Single-Variable Rules

The Single-Variable Rules tab allows you to create, view, and modify single-variable validation rules.

Rules. The list shows single-variable validation rules by name and the type of variable to which the rule can be applied. When the dialog box is opened, it shows rules defined in the data dictionary or, if no rules are currently defined, a placeholder rule called "Single-Variable Rule 1." The following buttons appear below the Rules list:

- **New.** Adds a new entry to the bottom of the Rules list. The rule is selected and assigned the name "SingleVarRule *n*," where *n* is an integer so that the new rule's name is unique among single-variable and cross-variable rules.

© Copyright IBM Corporation 1989, 2016

- **Duplicate.** Adds a copy of the selected rule to the bottom of the Rules list. The rule name is adjusted so that it is unique among single-variable and cross-variable rules. For example, if you duplicate "SingleVarRule 1," the name of the first duplicate rule would be "Copy of SingleVarRule 1," the second would be "Copy (2) of SingleVarRule 1," and so on.
- **Delete.** Deletes the selected rule.

Rule Definition. These controls allow you to view and set properties for a selected rule.

- **Name.** The name of the rule must be unique among single-variable and cross-variable rules.
- **Type.** This is the type of variable to which the rule can be applied. Select from **Numeric**, **String**, and **Date**.
- **Format.** This allows you to select the date format for rules that can be applied to date variables.
- **Valid Values.** You can specify the valid values either as a range or a list of values.

  Range Definition

  Range definition controls allow you to specify a valid range. Values outside the range are flagged as invalid.

  To specify a range, enter the minimum or maximum values, or both. The check box controls allow you to flag unlabeled and non-integer values within the range.

  List Definition

  List definition controls allow you to define a list of valid values. Values not included in the list are flagged as invalid.

  Enter list values in the grid. The check box determines whether case matters when string data values are checked against the list of acceptable values.
- **Allow user-missing values.** Controls whether user-missing values are flagged as invalid.
- **Allow system-missing values.** Controls whether system-missing values are flagged as invalid. This does not apply to string rule types.
- **Allow blank values.** Controls whether blank (that is, completely empty) string values are flagged as invalid. This does not apply to nonstring rule types.

Define Cross-Variable Rules

The Cross-Variable Rules tab allows you to create, view, and modify cross-variable validation rules.

Rules. The list shows cross-variable validation rules by name. When the dialog box is opened, it shows a placeholder rule called "CrossVarRule 1." The following buttons appear below the Rules list:

- **New.** Adds a new entry to the bottom of the Rules list. The rule is selected and assigned the name "CrossVarRule *n*," where *n* is an integer so that the new rule's name is unique among single-variable and cross-variable rules.
- **Duplicate.** Adds a copy of the selected rule to the bottom of the Rules list. The rule name is adjusted so that it is unique among single-variable and cross-variable rules. For example, if you duplicate "CrossVarRule 1," the name of the first duplicate rule would be "Copy of CrossVarRule 1," the second would be "Copy (2) of CrossVarRule 1," and so on.
- **Delete.** Deletes the selected rule.

Rule Definition. These controls allow you to view and set properties for a selected rule.

- **Name.** The name of the rule must be unique among single-variable and cross-variable rules.
- **Logical Expression.** This is, in essence, the rule definition. You should code the expression so that invalid cases evaluate to 1.

Building Expressions

1. To build an expression, either paste components into the Expression field or type directly in the Expression field.

- You can paste functions or commonly used system variables by selecting a group from the Function group list and double-clicking the function or variable in the Functions and Special Variables list (or select the function or variable and click **Insert**). Enter values for any parameters indicated by question marks (applies only to functions). The function group labeled **All** provides a list of all available functions and system variables. A brief description of the currently selected function or variable is displayed in a reserved area in the dialog box.
- String constants must be enclosed in quotation marks or apostrophes.
- If values contain decimals, a period (.) must be used as the decimal indicator.

Chapter 3. Validate Data

The Validate Data dialog box allows you to identify suspicious and invalid cases, variables, and data values in the active dataset.

Example. A data analyst must provide a monthly customer satisfaction report to her client. The data she receives every month needs to be quality checked for incomplete customer IDs, variable values that are out of range, and combinations of variable values that are commonly entered in error. The Validate Data dialog box allows the analyst to specify the variables that uniquely identify customers, define single-variable rules for the valid variable ranges, and define cross-variable rules to catch impossible combinations. The procedure returns a report of the problem cases and variables. Moreover, the data has the same data elements each month, so the analyst is able to apply the rules to the new data file next month.

Statistics. The procedure produces lists of variables, cases, and data values that fail various checks, counts of violations of single-variable and cross-variable rules, and simple descriptive summaries of analysis variables.

Weights. The procedure ignores the weight variable specification and instead treats it as any other analysis variable.

To Validate Data

1. From the menus choose:

   Data > **Validation** > **Validate Data...**
2. Select one or more analysis variables for validation by basic variable checks or by single-variable validation rules.

   Alternatively, you can:
3. Click the **Cross-Variable Rules** tab and apply one or more cross-variable rules.

Optionally, you can:

- Select one or more case identification variables to check for duplicate or incomplete IDs. Case ID variables are also used to label casewise output. If two or more case ID variables are specified, the combination of their values is treated as a case identifier.

Fields with Unknown Measurement Level

The Measurement Level alert is displayed when the measurement level for one or more variables (fields) in the dataset is unknown. Since measurement level affects the computation of results for this procedure, all variables must have a defined measurement level.

Scan Data. Reads the data in the active dataset and assigns default measurement level to any fields with a currently unknown measurement level. If the dataset is large, that may take some time.

Assign Manually. Opens a dialog that lists all fields with an unknown measurement level. You can use this dialog to assign measurement level to those fields. You can also assign measurement level in Variable View of the Data Editor.

Since measurement level is important for this procedure, you cannot access the dialog to run this procedure until all fields have a defined measurement level.

Validate Data Basic Checks

The Basic Checks tab allows you to select basic checks for analysis variables, case identifiers, and whole cases.

Analysis Variables. If you selected any analysis variables on the Variables tab, you can select any of the following checks of their validity. The check box allows you to turn the checks on or off.

- **Maximum percentage of missing values.** Reports analysis variables with a percentage of missing values greater than the specified value. The specified value must be a positive number less than or equal to 100.
- **Maximum percentage of cases in a single category.** If any analysis variables are categorical, this option reports categorical analysis variables with a percentage of cases representing a single nonmissing category greater than the specified value. The specified value must be a positive number less than or equal to 100. The percentage is based on cases with nonmissing values of the variable.
- **Maximum percentage of categories with count of 1.** If any analysis variables are categorical, this option reports categorical analysis variables in which the percentage of the variable's categories containing only one case is greater than the specified value. The specified value must be a positive number less than or equal to 100.
- **Minimum coefficient of variation.** If any analysis variables are scale, this option reports scale analysis variables in which the absolute value of the coefficient of variation is less than the specified value. This option applies only to variables in which the mean is nonzero. The specified value must be a non-negative number. Specifying 0 turns off the coefficient-of-variation check.
- **Minimum standard deviation.** If any analysis variables are scale, this option reports scale analysis variables whose standard deviation is less than the specified value. The specified value must be a non-negative number. Specifying 0 turns off the standard deviation check.

Case Identifiers. If you selected any case identifier variables on the Variables tab, you can select any of the following checks of their validity.

- **Flag incomplete IDs.** This option reports cases with incomplete case identifiers. For a particular case, an identifier is considered incomplete if the value of any ID variable is blank or missing.
- **Flag duplicate IDs.** This option reports cases with duplicate case identifiers. Incomplete identifiers are excluded from the set of possible duplicates.

Flag empty cases. This option reports cases in which all variables are empty or blank. For the purpose of identifying empty cases, you can choose to use all variables in the file (except any ID variables) or only analysis variables defined on the Variables tab.

Validate Data Single-Variable Rules

The Single-Variable Rules tab displays available single-variable validation rules and allows you to apply them to analysis variables. To define additional single-variable rules, click **Define Rules**. See the topic "Define Single-Variable Rules" on page 3 for more information.

Analysis Variables. The list shows analysis variables, summarizes their distributions, and shows the number of rules applied to each variable. Note that user- and system-missing values are not included in the summaries. The Display drop-down list controls which variables are shown; you can choose from **All variables**, **Numeric variables**, **String variables**, and **Date variables**.

Rules. To apply rules to analysis variables, select one or more variables and check all rules that you want to apply in the Rules list. The Rules list shows only rules that are appropriate for the selected analysis variables. For example, if numeric analysis variables are selected, only numeric rules are shown; if a string variable is selected, only string rules are shown. If no analysis variables are selected or they have mixed data types, no rules are shown.

Variable Distributions. The distribution summaries shown in the Analysis Variables list can be based on all cases or on a scan of the first *n* cases, as specified in the Cases text box. Clicking **Rescan** updates the distribution summaries.

Validate Data Cross-Variable Rules

The Cross-Variable Rules tab displays available cross-variable rules and allows you to apply them to your data. To define additional cross-variable rules, click **Define Rules**. See the topic "Define Cross-Variable Rules" on page 4 for more information.

Validate Data Output

Casewise Report. If you have applied any single-variable or cross-variable validation rules, you can request a report that lists validation rule violations for individual cases.

- **Minimum Number of Violations.** This option specifies the minimum number of rule violations required for a case to be included in the report. Specify a positive integer.
- **Maximum Number of Cases.** This option specifies the maximum number of cases included in the case report. Specify a positive integer less than or equal to 1000.

Single-Variable Validation Rules. If you have applied any single-variable validation rules, you can choose how to display the results or whether to display them at all.

- **Summarize violations by analysis variable.** For each analysis variable, this option shows all single-variable validation rules that were violated and the number of values that violated each rule. It also reports the total number of single-variable rule violations for each variable.
- **Summarize violations by rule.** For each single-variable validation rule, this option reports variables that violated the rule and the number of invalid values per variable. It also reports the total number of values that violated each rule across variables.

Display descriptive statistics for analysis variables. This option allows you to request descriptive statistics for analysis variables. A frequency table is generated for each categorical variable. A table of summary statistics including the mean, standard deviation, minimum, and maximum is generated for the scale variables.

Move cases with validation rule violations to the top of the active dataset. This option moves cases with single-variable or cross-variable rule violations to the top of the active dataset for easy perusal.

Validate Data Save

The Save tab allows you to save variables that record rule violations to the active dataset.

Summary Variables. These are individual variables that can be saved. Check a box to save the variable. Default names for the variables are provided; you can edit them.

- **Empty case indicator.** Empty cases are assigned the value 1. All other cases are coded 0. Values of the variable reflect the scope specified on the Basic Checks tab.
- **Duplicate ID Group.** Cases that have the same case identifier (other than cases with incomplete identifiers) are assigned the same group number. Cases with unique or incomplete identifiers are coded 0.
- **Incomplete ID indicator.** Cases with empty or incomplete case identifiers are assigned the value 1. All other cases are coded 0.
- **Validation rule violations.** This is the casewise total count of single-variable and cross-variable validation rule violations.

Replace existing summary variables. Variables saved to the data file must have unique names or replace variables with the same name.

Save indicator variables. This option allows you to save a complete record of validation rule violations. Each variable corresponds to an application of a validation rule and has a value of 1 if the case violates the rule and a value of 0 if it does not.

Chapter 4. Automated Data Preparation

Preparing data for analysis is one of the most important steps in any project—and traditionally, one of the most time consuming. Automated Data Preparation (ADP) handles the task for you, analyzing your data and identifying fixes, screening out fields that are problematic or not likely to be useful, deriving new attributes when appropriate, and improving performance through intelligent screening techniques. You can use the algorithm in fully **automatic** fashion, allowing it to choose and apply fixes, or you can use it in **interactive** fashion, previewing the changes before they are made and accept or reject them as you want.

Using ADP enables you to make your data ready for model building quickly and easily, without needing prior knowledge of the statistical concepts involved. Models will tend to build and score more quickly; in addition, using ADP improves the robustness of automated modeling processes.

Note: when ADP prepares a field for analysis, it creates a new field containing the adjustments or transformations, rather than replacing the existing values and properties of the old field. The old field is not used in further analysis; its role is set to None. Also note that any user-missing value information is not transferred to these newly created fields, and any missing values in the new field are system-missing.

Example. An insurance company with limited resources to investigate homeowner's insurance claims wants to build a model for flagging suspicious, potentially fraudulent claims. Before building the model, they will ready the data for modeling using automated data preparation. Since they want to be able to review the proposed transformations before the transformations are applied, they will use automated data preparation in interactive mode.

An automotive industry group keeps track of the sales for a variety of personal motor vehicles. In an effort to be able to identify over- and underperforming models, they want to establish a relationship between vehicle sales and vehicle characteristics. They will use automated data preparation to prepare the data for analysis, and build models using the data "before" and "after" preparation to see how the results differ.

What is your objective? Automated data preparation recommends data preparation steps that will affect the speed with which other algorithms can build models and improve the predictive power of those models. This can include transforming, constructing and selecting features. The target can also be transformed. You can specify the model-building priorities that the data preparation process should concentrate on.

- **Balance speed and accuracy.** This option prepares the data to give equal priority to both the speed with which data are processed by model-building algorithms and the accuracy of the predictions.
- **Optimize for speed.** This option prepares the data to give priority to the speed with which data are processed by model-building algorithms. When you are working with very large datasets, or are looking for a quick answer, select this option.
- **Optimize for accuracy.** This option prepares the data to give priority to the accuracy of predictions produced by model-building algorithms.
- **Custom analysis.** When you want to manually change the algorithm on the Settings tab, select this option. Note that this setting is automatically selected if you subsequently make changes to options on the Settings tab that are incompatible with one of the other objectives.

© Copyright IBM Corporation 1989, 2016

To Obtain Automatic Data Preparation

From the menus choose:

1. From the menus choose:
   Transform > **Prepare Data for Modeling** > **Automatic...**
2. Click **Run**.

Optionally, you can:

- Specify an objective on the Objective tab.
- Specify field assignments on the Fields tab.
- Specify expert settings on the Settings tab.

To Obtain Interactive Data Preparation

1. From the menus choose:
   Transform > **Prepare Data for Modeling** > **Interactive...**
2. Click **Analyze** in the toolbar at the top of the dialog.
3. Click the Analysis tab and review the suggested data preparation steps.
4. If satisfied, click **Run**. Otherwise, click **Clear Analysis**, change any settings you want, and click **Analyze**.

Optionally, you can:

- Specify an objective on the Objective tab.
- Specify field assignments on the Fields tab.
- Specify expert settings on the Settings tab.
- Save the suggested data preparation steps to an XML file by clicking **Save XML**.

Fields Tab

The Fields tab specifies which fields should be prepared for further analysis.

Use predefined roles. This option uses existing field information. If there is a single field with a role as a Target, it will be used as the target; otherwise there will be no target. All fields with a predefined role as an Input will be used as inputs. At least one input field is required.

Use custom field assignments. When you override field roles by moving fields from their default lists, the dialog automatically switches to this option. When making custom field assignments, specify the following fields:

- **Target (optional).** If you plan to build models that require a target, select the target field. This is similar to setting the field role to Target.
- **Inputs.** Select one or more input fields. This is similar to setting the field role to Input.

Settings Tab

The Settings tab comprises several different groups of settings that you can modify to fine-tune how the algorithm processes your data. If you make any changes to the default settings that are incompatible with the other objectives, the Objective tab is automatically updated to select the **Customize analysis** option.

Prepare Dates & Times

Many modeling algorithms are unable to directly handle date and time details; these settings enable you to derive new duration data that can be used as model inputs from dates and times in your existing data. The fields containing dates and times must be predefined with date or time storage types. The original date and time fields will not be recommended as model inputs following automated data preparation.

Prepare dates and times for modeling. Deselecting this option disables all other Prepare Dates & Times controls while maintaining the selections.

Compute elapsed time until reference date. This produces the number of years/months/days since a reference date for each variable containing dates.

- **Reference Date.** Specify the date from which the duration will be calculated with regard to the date information in the input data. Selecting **Today's date** means that the current system date is always used when ADP is executed. To use a specific date, select **Fixed date** and enter the required date.
- **Units for Date Duration.** Specify whether ADP should automatically decide on the date duration unit, or select from **Fixed units** of Years, Months, or Days.

Compute elapsed time until reference time. This produces the number of hours/minutes/seconds since a reference time for each variable containing times.

- **Reference Time.** Specify the time from which the duration will be calculated with regard to the time information in the input data. Selecting **Current time** means that the current system time is always used when ADP is executed. To use a specific time, select **Fixed time** and enter the required details.
- **Units for Time Duration.** Specify whether ADP should automatically decide on the time duration unit, or select from **Fixed units** of Hours, Minutes, or Seconds.

Extract Cyclical Time Elements. Use these settings to split a single date or time field into one or more fields. For example if you select all three date checkboxes, the input date field "1954-05-23" is split into three fields: 1954, 5, and 23, each using the suffix defined on the **Field Names** panel, and the original date field is ignored.

- **Extract from dates.** For any date inputs, specify if you want to extract years, months, days, or any combination.
- **Extract from times.** For any time inputs, specify if you want to extract hours, minutes, seconds, or any combination.

Exclude Fields

Poor quality data can affect the accuracy of your predictions; therefore, you can specify the acceptable quality level for input features. All fields that are constant or have 100% missing values are automatically excluded.

Exclude low quality input fields. Deselecting this option disables all other Exclude Fields controls while maintaining the selections.

Exclude fields with too many missing values. Fields with more than the specified percentage of missing values are removed from further analysis. Specify a value greater than or equal to 0, which is equivalent to deselecting this option, and less than or equal to 100, though fields with all missing values are automatically excluded. The default is 50.

Exclude nominal fields with too many unique categories. Nominal fields with more than the specified number of categories are removed from further analysis. Specify a positive integer. The default is 100. This is useful for automatically removing fields containing record-unique information from modeling, like ID, address, or name.

Exclude categorical fields with too many values in a single category. Ordinal and nominal fields with a category that contains more than the specified percentage of the records are removed from further analysis. Specify a value greater than or equal to 0, equivalent to deselecting this option, and less than or equal to 100, though constant fields are automatically excluded. The default is 95.

Adjust Measurement

Adjust measurement level. Deselecting this option disables all other Adjust Measurement controls while maintaining the selections.

Measurement Level. Specify whether the measurement level of continuous fields with "too few" values can be adjusted to ordinal, and ordinal fields with "too many" values can be adjusted to continuous.

- **Maximum number of values for ordinal fields.** Ordinal fields with more than the specified number of categories are recast as continuous fields. Specify a positive integer. The default is 10. This value must be greater than or equal to the minimum number of values for continuous fields.
- **Minimum number of values for continuous fields.** Continuous fields with less than the specified number of unique values are recast as ordinal fields. Specify a positive integer. The default is 5. This value must be less than or equal to the maximum number of values for ordinal fields.

Improve Data Quality

Prepare fields to improve data quality. Deselecting this option disables all other Improve Data Quality controls while maintaining the selections.

Outlier Handling. Specify whether to replace outliers for the inputs and target; if so, specify an outlier cutoff criterion, measured in standard deviations, and a method for replacing outliers. Outliers can be replaced by either trimming (setting to the cutoff value), or by setting them as missing values. Any outliers set to missing values follow the missing value handling settings selected below.

Replace Missing Values. Specify whether to replace missing values of continuous, nominal, or ordinal fields.

Reorder Nominal Fields. Select this to recode the values of nominal (set) fields from smallest (least frequently occurring) to largest (most frequently occurring) category. The new field values start with 0 as the least frequent category. Note that the new field will be numeric even if the original field is a string. For example, if a nominal field's data values are "A", "A", "A", "B", "C", "C", then automated data preparation would recode "B" into 0, "C" into 1, and "A" into 2.

Rescale Fields

Rescale fields. Deselecting this option disables all other Rescale Fields controls while maintaining the selections.

Analysis Weight. This variable contains analysis (regression or sampling) weights. Analysis weights are used to account for differences in variance across levels of the target field. Select a continuous field.

Continuous Input Fields. This will normalize continuous input fields using a **z-score transformation** or **min/max transformation**. Rescaling inputs is especially useful when you select **Perform feature construction** on the Select and Construct settings.

- **Z-score transformation.** Using the observed mean and standard deviation as population parameter estimates, the fields are standardized and then the *z* scores are mapped to the corresponding values of a normal distribution with the specified **Final mean** and **Final standard deviation**. Specify a number for **Final mean** and a positive number for **Final standard deviation**. The defaults are 0 and 1, respectively, corresponding to standardized rescaling.

- **Min/max transformation.** Using the observed minimum and maximum as population parameter estimates, the fields are mapped to the corresponding values of a uniform distribution with the specified **Minimum** and **Maximum**. Specify numbers with **Maximum** greater than **Minimum**.

Continuous Target. This transforms a continuous target using the Box-Cox transformation into a field that has an approximately normal distribution with the specified **Final mean** and **Final standard deviation**. Specify a number for **Final mean** and a positive number for **Final standard deviation**. The defaults are 0 and 1, respectively.

Note: If a target has been transformed by ADP, subsequent models built using the transformed target score the transformed units. In order to interpret and use the results, you must convert the predicted value back to the original scale. See the topic for more information. See the topic "Backtransform Scores" on page 23 for more information.

Transform Fields

To improve the predictive power of your data, you can transform the input fields.

Transform field for modeling. Deselecting this option disables all other Transform Fields controls while maintaining the selections.

Categorical Input Fields The following options are available:

- **Merge sparse categories to maximize association with target.** Select this to make a more parsimonious model by reducing the number of fields to be processed in association with the target. Similar categories are identified based upon the relationship between the input and the target. Categories that are not significantly different (that is, having a *p*-value greater than the value specified) are merged. Specify a value greater than 0 and less than or equal to 1. If all categories are merged into one, the original and derived versions of the field are excluded from further analysis because they have no value as a predictor.
- **When there is no target, merge sparse categories based on counts.** If the dataset has no target, you can choose to merge sparse categories of ordinal and nominal fields. The equal frequency method is used to merge categories with less than the specified minimum percentage of the total number of records. Specify a value greater than or equal to 0 and less than or equal to 100. The default is 10. Merging stops when there are not categories with less than the specified minimum percent of cases, or when there are only two categories left.

Continuous Input Fields. If the dataset includes a categorical target, you can bin continuous inputs with strong associations to improve processing performance. Bins are created based upon the properties of "homogeneous subsets", which are identified by the Scheffe method using the specified *p*-value as the alpha for the critical value for determining homogeneous subsets. Specify a value greater than 0 and less than or equal to 1. The default is 0.05. If the binning operation results in a single bin for a particular field, the original and binned versions of the field are excluded because they have no value as a predictor.

Note: Binning in ADP differs from optimal binning. Optimal binning uses entropy information to convert a continuous field to a categorical field; this needs to sort data and store it all in memory. ADP uses homogeneous subsets to bin a continuous field, which means that ADP binning does not need to sort data and does not store all data in memory. The use of the homogeneous subset method to bin a continuous field means that the number of categories after binning is always less than or equal to the number of categories in the target.

Select and Construct

To improve the predictive power of your data, you can construct new fields based on the existing fields.

Perform feature selection. A continuous input is removed from the analysis if the *p*-value for its correlation with the target is greater than the specified *p*-value.

Perform feature construction. Select this option to derive new features from a combination of several existing features. The old features are not used in further analysis. This option only applies to continuous input features where the target is continuous, or where there is no target.

Field Names

To easily identify new and transformed features, ADP creates and applies basic new names, prefixes, or suffixes. You can amend these names to be more relevant to your own needs and data.

Transformed and Constructed Fields. Specify the name extensions to be applied to transformed target and input fields.

In addition, specify the prefix name to be applied to any features that are constructed via the Select and Construct settings. The new name is created by attaching a numeric suffix to this prefix root name. The format of the number depends on how many new features are derived, for example:

- 1-9 constructed features will be named: feature1 to feature9.
- 10-99 constructed features will be named: feature01 to feature99.
- 100-999 constructed features will be named: feature001 to feature999, and so on.

This ensures that the constructed features will sort in a sensible order no matter how many there are.

Durations Computed from Dates and Times. Specify the name extensions to be applied to durations computed from both dates and times.

Cyclical Elements Extracted from Dates and Times. Specify the name extensions to be applied to cyclical elements extracted from both dates and times.

Applying and Saving Transformations

Depending upon whether you are using the Interactive or Automatic Data Preparation dialogs, the settings for applying and saving transformations are slightly different.

Interactive Data Preparation Apply Transformations Settings

Transformed Data. These settings specify where to save the transformed data.

- **Add new fields to the active dataset.** Any fields created by automated data preparation are added as new fields to the active dataset. **Update roles for analyzed fields** will set the role to None for any fields that are excluded from further analysis by automated data preparation.
- **Create a new dataset or file containing the transformed data.** Fields recommended by automated data preparation are added to a new dataset or file. **Include unanalyzed fields** adds fields in the original dataset that were not specified on the Fields tab to the new dataset. This is useful for transferring fields containing information not used in modeling, like ID or address, or name, into the new dataset.

Automatic Data Preparation Apply and Save Settings

The Transformed Data group is the same as in Interactive Data Preparation. In Automatic Data preparation, the following additional options are available:

Apply transformations. In the Automatic Data Preparation dialogs, deselecting this option disables all other Apply and Save controls while maintaining the selections.

Save transformations as syntax. This saves the recommended transformations as command syntax to an external file. The Interactive Data Preparation dialog does not have this control because it will paste the transformations as command syntax to the syntax window if you click **Paste**.

Save transformations as XML. This saves the recommended transformations as XML to an external file, which can be merged with model PMML using `TMS MERGE` or applied to another dataset using `TMS IMPORT`. The Interactive Data Preparation dialog does not have this control because it will save the transformations as XML if you click **Save XML** in the toolbar at the top of the dialog.

Analysis Tab

Note: The Analysis tab is used in the Interactive Data Preparation dialog to allow you to review the recommended transformations. The Automatic Data Preparation dialog does not include this step.

1. When you are satisfied with the ADP settings, including any changes made on the Objective, Fields, and Settings tabs, click **Analyze Data**; the algorithm applies the settings to the data inputs and displays the results on the Analysis tab.

The Analysis tab contains both tabular and graphical output that summarizes the processing of your data and displays recommendations as to how the data may be modified or improved for scoring. You can then review and either accept or reject those recommendations.

The Analysis tab is made up of two panels, the main view on the left and the linked, or auxiliary, view on the right. There are three main views:

- Field Processing Summary (the default). See the topic "Field Processing Summary" for more information.
- Fields. See the topic "Fields" on page 18 for more information.
- Action Summary. See the topic "Action Summary" on page 19 for more information.

There are four linked/auxiliary views:

- Predictive Power (the default). See the topic "Predictive Power" on page 19 for more information.
- Fields Table. See the topic "Fields Table" on page 19 for more information.
- Field Details. See the topic "Field Details" on page 20 for more information.
- Action Details. See the topic "Action Details" on page 21 for more information.

Links between views

Within the main view, underlined text in the tables controls the display in the linked view. Clicking on the text allows you to get details on a particular field, set of fields, or processing step. The link that you last selected is shown in a darker color; this helps you identify the connection between the contents of the two view panels.

Resetting the views

To redisplay the original Analysis recommendations and abandon any changes you have made to the Analysis views, click **Reset** at the bottom of the main view panel.

Field Processing Summary

The Field Processing Summary table gives a snapshot of the projected overall impact of processing, including changes to the state of the features and the number of features constructed.

Note that no model is actually built, so there isn't a measure or graph of the change in overall predictive power before and after data preparation; instead, you can display graphs of the predictive power of individual recommended predictors.

The table displays the following information:

- The number of target fields.

- The number of original (input) predictors.
- The predictors recommended for use in analysis and modeling. This includes the total number of fields recommended; the number of original, untransformed, fields recommended; the number of transformed fields recommended (excluding intermediate versions of any field, fields derived from date/time predictors, and constructed predictors); the number of fields recommended that are derived from date/time fields; and the number of constructed predictors recommended.
- The number of input predictors not recommended for use in any form, whether in their original form, as a derived field, or as input to a constructed predictor.

Where any of the **Fields** information is underlined, click to display more details in a linked view. Details of the **Target**, **Input features**, and **Input features not used** are shown in the Fields Table linked view. See the topic "Fields Table" on page 19 for more information. **Features recommended for use in analysis** are displayed in the Predictive Power linked view. See the topic "Predictive Power" on page 19 for more information.

Fields

The Fields main view displays the processed fields and whether ADP recommends using them in downstream models. You can override the recommendation for any field; for example, to exclude constructed features or include features that ADP recommends excluding. If a field has been transformed, you can decide whether to accept the suggested transformation or use the original version.

The Fields view consists of two tables, one for the target and one for predictors that were either processed or created.

Target table

The **Target** table is only shown if a target is defined in the data.

The table contains two columns:

- **Name.** This is the name or label of the target field; the original name is always used, even if the field has been transformed.
- **Measurement Level.** This displays the icon representing the measurement level; hover the mouse over the icon to display a label (continuous, ordinal, nominal, and so on) that describes the data.

  If the target has been transformed the **Measurement Level** column reflects the final transformed version. *Note*: you cannot turn off transformations for the target.

Predictors table

The **Predictors** table is always shown. Each row of the table represents a field. By default the rows are sorted in descending order of predictive power.

For ordinary features, the original name is always used as the row name. Both original and derived versions of date/time fields appear in the table (in separate rows); the table also includes constructed predictors.

Note that transformed versions of fields shown in the table always represent the final versions.

By default only recommended fields are shown in the Predictors table. To display the remaining fields, select the **Include nonrecommended fields in table** box above the table; these fields are then displayed at the bottom of the table.

The table contains the following columns:

- **Version to Use.** This displays a drop-down list that controls whether a field will be used downstream and whether to use the suggested transformations. By default, the drop-down list reflects the recommendations.

  For ordinary predictors that have been transformed the drop-down list has three choices: **Transformed**, **Original**, and **Do not use**.

  For untransformed ordinary predictors the choices are: **Original** and **Do not use**.

  For derived date/time fields and constructed predictors the choices are: **Transformed** and **Do not use**.

  For original date fields the drop-down list is disabled and set to **Do not use**.

  Note: For predictors with both original and transformed versions, changing between **Original** and **Transformed** versions automatically updates the **Measurement Level** and **Predictive Power** settings for those features.
- **Name.** Each field's name is a link. Click a name to display more information about the field in the linked view. See the topic "Field Details" on page 20 for more information.
- **Measurement Level.** This displays the icon representing the data type; hover the mouse over the icon to display a label (continuous, ordinal, nominal, and so on) that describes the data.
- **Predictive Power.** Predictive power is displayed only for fields that ADP recommends. This column is not displayed if there is no target defined. Predictive power ranges from 0 to 1, with larger values indicating "better" predictors. In general, predictive power is useful for comparing predictors within an ADP analysis, but predictive power values should not be compared across analyses.

Action Summary

For each action taken by automated data preparation, input predictors are transformed and/or filtered out; fields that survive one action are used in the next. The fields that survive through to the last step are then recommended for use in modeling, whilst inputs to transformed and constructed predictors are filtered out.

The Action Summary is a simple table that lists the processing actions taken by ADP. Where any **Action** is underlined, click to display more details in a linked view about the actions taken. See the topic "Action Details" on page 21 for more information.

Note: Only the original and final transformed versions of each field are shown, not any intermediate versions that were used during analysis.

Predictive Power

Displayed by default when the analysis is first run, or when you select **Predictors recommended for use in analysis** in the Field Processing Summary main view, the chart displays the predictive power of recommended predictors. Fields are sorted by predictive power, with the field with the highest value appearing at the top.

For transformed versions of ordinary predictors, the field name reflects your choice of suffix in the Field Names panel of the Settings tab; for example: *_transformed*.

Measurement level icons are displayed after the individual field names.

The predictive power of each recommended predictor is computed from either a linear regression or naïve Bayes model, depending upon whether the target is continuous or categorical.

Fields Table

Displayed when you click **Target**, **Predictors**, or **Predictors not used** in the Field Processing Summary main view, the Fields Table view displays a simple table listing the relevant features.

The table contains two columns:

- **Name.** The predictor name.

  For targets, the original name or label of the field is used, even if the target has been transformed.

  For transformed versions of ordinary predictors, the name reflects your choice of suffix in the Field Names panel of the Settings tab; for example: *_transformed*.

  For fields derived from dates and times, the name of the final transformed version is used; for example: *bdate_years*.

  For constructed predictors, the name of the constructed predictor is used; for example: *Predictor1*.
- **Measurement Level.** This displays the icon representing the data type.

  For the Target, the **Measurement Level** always reflects the transformed version (if the target has been transformed); for example, changed from ordinal (ordered set) to continuous (range, scale), or vice versa.

Field Details

Displayed when you click any **Name** in the Fields main view, the Field Details view contains distribution, missing values, and predictive power charts (if applicable) for the selected field. In addition, the processing history for the field and the name of the transformed field are also shown (if applicable).

For each chart set, two versions are shown side by side to compare the field with and without transformations applied; if a transformed version of the field does not exist, a chart is shown for the original version only. For derived date or time fields and constructed predictors, the charts are only shown for the new predictor.

Note: If a field is excluded due to having too many categories only the processing history is shown.

Distribution Chart

Continuous field distribution is shown as a histogram, with a normal curve overlaid, and a vertical reference line for the mean value; categorical fields are displayed as a bar chart.

Histograms are labeled to show standard deviation and skewness; however, skewness is not displayed if the number of values is 2 or fewer or the variance of the original field is less than 10-20.

Hover the mouse over the chart to display either the mean for histograms, or the count and percentage of the total number of records for categories in bar charts.

Missing Value Chart

Pie charts compare the percentage of missing values with and without transformations applied; the chart labels show the percentage.

If ADP carried out missing value handling, the post-transformation pie chart also includes the replacement value as a label -- that is, the value used in place of missing values.

Hover the mouse over the chart to display the missing value count and percentage of the total number of records.

Predictive Power Chart

For recommended fields, bar charts display the predictive power before and after transformation. If the target has been transformed, the calculated predictive power is in respect to the transformed target.

Note: Predictive power charts are not shown if no target is defined, or if the target is clicked in the main view panel.

Hover the mouse over the chart to display the predictive power value.

Processing History Table

The table shows how the transformed version of a field was derived. Actions taken by ADP are listed in the order in which they were carried out; however, for certain steps multiple actions may have been carried out for a particular field.

Note: This table is not shown for fields that have not been transformed.

The information in the table is broken down into two or three columns:

- **Action.** The name of the action. For example, Continuous Predictors. See the topic "Action Details" for more information.
- **Details.** The list of processing carried out. For example, Transform to standard units.
- **Function.** Only shown for constructed predictors, this displays the linear combination of input fields, for example, .06*age + 1.21*height.

Action Details

Displayed when you select any underlined **Action** in the Action Summary main view, the Action Details linked view displays both action-specific and common information for each processing step that was carried out; the action-specific details are displayed first.

For each action, the description is used as the title at the top of the linked view. The action-specific details are displayed below the title, and may include details of the number of derived predictors, fields recast, target transformations, categories merged or reordered, and predictors constructed or excluded.

As each action is processed, the number of predictors used in the processing may change, for example as predictors are excluded or merged.

Note: If an action was turned off, or no target was specified, an error message is displayed in place of the action details when the action is clicked in the Action Summary main view.

There are nine possible actions; however, not all are necessarily active for every analysis.

Text Fields Table

The table displays the number of:

- Predictors excluded from analysis.

Date and Time Predictors Table

The table displays the number of:

- Durations derived from date and time predictors.
- Date and time elements.
- Derived date and time predictors, in total.

The reference date or time is displayed as a footnote if any date durations were calculated.

Predictor Screening Table

The table displays the number of the following predictors excluded from processing:

- Constants.
- Predictors with too many missing values.

- Predictors with too many cases in a single category.
- Nominal fields (sets) with too many categories.
- Predictors screened out, in total.

Check Measurement Level Table

The table displays the numbers of fields recast, broken down into the following:
- Ordinal fields (ordered sets) recast as continuous fields.
- Continuous fields recast as ordinal fields.
- Total number recast.

If no input fields (target or predictors) were continuous or ordinal, this is shown as a footnote.

Outliers Table

The table displays counts of how any outliers were handled.
- Either the number of continuous fields for which outliers were found and trimmed, or the number of continuous fields for which outliers were found and set to missing, depending on your settings in the Prepare Inputs & Target panel on the Settings tab.
- The number of continuous fields excluded because they were constant, after outlier handling.

One footnote shows the outlier cutoff value; while another footnote is shown if no input fields (target or predictors) were continuous.

Missing Values Table

The table displays the numbers of fields that had missing values replaced, broken down into:
- Target. This row is not shown if no target is specified.
- Predictors. This is further broken down into the number of nominal (set), ordinal (ordered set), and continuous.
- The total number of missing values replaced.

Target Table

The table displays whether the target was transformed, shown as:
- Box-Cox transformation to normality. This is further broken down into columns that show the specified criteria (mean and standard deviation) and Lambda.
- Target categories reordered to improve stability.

Categorical Predictors Table

The table displays the number of categorical predictors:
- Whose categories were reordered from lowest to highest to improve stability.
- Whose categories were merged to maximize association with the target.
- Whose categories were merged to handle sparse categories.
- Excluded due to low association with the target.
- Excluded because they were constant after merging.

A footnote is shown if there were no categorical predictors.

Continuous Predictors Table

There are two tables. The first displays one of the following number of transformations:

- Predictor values transformed to standard units. In addition, this shows the number of predictors transformed, the specified mean, and the standard deviation.
- Predictor values mapped to a common range. In addition, this shows the number of predictors transformed using a min-max transformation, as well as the specified minimum and maximum values.
- Predictor values binned and the number of predictors binned.

The second table displays the predictor space construction details, shown as the number of predictors:

- Constructed.
- Excluded due to a low association with the target.
- Excluded because they were constant after binning.
- Excluded because they were constant after construction.

A footnote is shown if no continuous predictors were input.

Backtransform Scores

If a target has been transformed by ADP, subsequent models built using the transformed target score the transformed units. In order to interpret and use the results, you must convert the predicted value back to the original scale.

1. To backtransform scores, from the menus choose:

   Transform > **Prepare Data for Modeling** > **Backtransform Scores...**
2. Select a field to backtransform. This field should contain model-predicted values of the transformed target.
3. Specify a suffix for the new field. This new field will contain model-predicted values in the original scale of the untransformed target.
4. Specify the location of the XML file containing the ADP transformations. This should be a file saved from the Interactive or Automatic Data Preparation dialogs. See the topic "Applying and Saving Transformations" on page 16 for more information.

Chapter 5. Identify Unusual Cases

The Anomaly Detection procedure searches for unusual cases based on deviations from the norms of their cluster groups. The procedure is designed to quickly detect unusual cases for data-auditing purposes in the exploratory data analysis step, prior to any inferential data analysis. This algorithm is designed for generic anomaly detection; that is, the definition of an anomalous case is not specific to any particular application, such as detection of unusual payment patterns in the healthcare industry or detection of money laundering in the finance industry, in which the definition of an anomaly can be well-defined.

Example. A data analyst hired to build predictive models for stroke treatment outcomes is concerned about data quality because such models can be sensitive to unusual observations. Some of these outlying observations represent truly unique cases and are thus unsuitable for prediction, while other observations are caused by data entry errors in which the values are technically "correct" and thus cannot be caught by data validation procedures. The Identify Unusual Cases procedure finds and reports these outliers so that the analyst can decide how to handle them.

Statistics. The procedure produces peer groups, peer group norms for continuous and categorical variables, anomaly indices based on deviations from peer group norms, and variable impact values for variables that most contribute to a case being considered unusual.

Data Considerations

Data. This procedure works with both continuous and categorical variables. Each row represents a distinct observation, and each column represents a distinct variable upon which the peer groups are based. A case identification variable can be available in the data file for marking output, but it will not be used in the analysis. Missing values are allowed. The weight variable, if specified, is ignored.

The detection model can be applied to a new test data file. The elements of the test data must be the same as the elements of the training data. And, depending on the algorithm settings, the missing value handling that is used to create the model may be applied to the test data file prior to scoring.

Case order. Note that the solution may depend on the order of cases. To minimize order effects, randomly order the cases. To verify the stability of a given solution, you may want to obtain several different solutions with cases sorted in different random orders. In situations with extremely large file sizes, multiple runs can be performed with a sample of cases sorted in different random orders.

Assumptions. The algorithm assumes that all variables are nonconstant and independent and that no case has missing values for any of the input variables. Each continuous variable is assumed to have a normal (Gaussian) distribution, and each categorical variable is assumed to have a multinomial distribution. Empirical internal testing indicates that the procedure is fairly robust to violations of both the assumption of independence and the distributional assumptions, but be aware of how well these assumptions are met.

To Identify Unusual Cases

1. From the menus choose:
   Data > **Identify Unusual Cases...**
2. Select at least one analysis variable.
3. Optionally, choose a case identifier variable to use in labeling output.

Fields with Unknown Measurement Level

The Measurement Level alert is displayed when the measurement level for one or more variables (fields) in the dataset is unknown. Since measurement level affects the computation of results for this procedure, all variables must have a defined measurement level.

Scan Data. Reads the data in the active dataset and assigns default measurement level to any fields with a currently unknown measurement level. If the dataset is large, that may take some time.

Assign Manually. Opens a dialog that lists all fields with an unknown measurement level. You can use this dialog to assign measurement level to those fields. You can also assign measurement level in Variable View of the Data Editor.

Since measurement level is important for this procedure, you cannot access the dialog to run this procedure until all fields have a defined measurement level.

Identify Unusual Cases Output

List of unusual cases and reasons why they are considered unusual. This option produces three tables:

- The anomaly case index list displays cases that are identified as unusual and displays their corresponding anomaly index values.
- The anomaly case peer ID list displays unusual cases and information concerning their corresponding peer groups.
- The anomaly reason list displays the case number, the reason variable, the variable impact value, the value of the variable, and the norm of the variable for each reason.

All tables are sorted by anomaly index in descending order. Moreover, the IDs of the cases are displayed if the case identifier variable is specified on the Variables tab.

Summaries. The controls in this group produce distribution summaries.

- **Peer group norms.** This option displays the continuous variable norms table (if any continuous variable is used in the analysis) and the categorical variable norms table (if any categorical variable is used in the analysis). The continuous variable norms table displays the mean and standard deviation of each continuous variable for each peer group. The categorical variable norms table displays the mode (most popular category), frequency, and frequency percentage of each categorical variable for each peer group. The mean of a continuous variable and the mode of a categorical variable are used as the norm values in the analysis.
- **Anomaly indices.** The anomaly index summary displays descriptive statistics for the anomaly index of the cases that are identified as the most unusual.
- **Reason occurrence by analysis variable.** For each reason, the table displays the frequency and frequency percentage of each variable's occurrence as a reason. The table also reports the descriptive statistics of the impact of each variable. If the maximum number of reasons is set to 0 on the Options tab, this option is not available.
- **Cases processed.** The case processing summary displays the counts and count percentages for all cases in the active dataset, the cases included and excluded in the analysis, and the cases in each peer group.

Identify Unusual Cases Save

Save Variables. Controls in this group allow you to save model variables to the active dataset. You can also choose to replace existing variables whose names conflict with the variables to be saved.

- **Anomaly index.** Saves the value of the anomaly index for each case to a variable with the specified name.
- **Peer groups.** Saves the peer group ID, case count, and size as a percentage for each case to variables with the specified rootname. For example, if the rootname *Peer* is specified, the variables *Peerid*, *PeerSize*, and *PeerPctSize* are generated. *Peerid* is the peer group ID of the case, *PeerSize* is the group's size, and *PeerPctSize* is the group's size as a percentage.

- **Reasons.** Saves sets of reasoning variables with the specified rootname. A set of reasoning variables consists of the name of the variable as the reason, its variable impact measure, its own value, and the norm value. The number of sets depends on the number of reasons requested on the Options tab. For example, if the rootname *Reason* is specified, the variables *ReasonVar_k*, *ReasonMeasure_k*, *ReasonValue_k*, and *ReasonNorm_k* are generated, where *k* is the *k*th reason. This option is not available if the number of reasons is set to 0.

Export Model File. Allows you to save the model in XML format.

Identify Unusual Cases Missing Values

The Missing Values tab is used to control handling of user-missing and system-missing values.

- **Exclude missing values from analysis.** Cases with missing values are excluded from the analysis.
- **Include missing values in analysis.** Missing values of continuous variables are substituted with their corresponding grand means, and missing categories of categorical variables are grouped and treated as a valid category. The processed variables are then used in the analysis. Optionally, you can request the creation of an additional variable that represents the proportion of missing variables in each case and use that variable in the analysis.

Identify Unusual Cases Options

Criteria for Identifying Unusual Cases. These selections determine how many cases are included in the anomaly list.

- **Percentage of cases with highest anomaly index values.** Specify a positive number that is less than or equal to 100.
- **Fixed number of cases with highest anomaly index values.** Specify a positive integer that is less than or equal to the total number of cases in the active dataset that are used in the analysis.
- **Identify only cases whose anomaly index value meets or exceeds a minimum value.** Specify a non-negative number. A case is considered anomalous if its anomaly index value is larger than or equal to the specified cutoff point. This option is used together with the **Percentage of cases** and **Fixed number of cases** options. For example, if you specify a fixed number of 50 cases and a cutoff value of 2, the anomaly list will consist of, at most, 50 cases, each with an anomaly index value that is larger than or equal to 2.

Number of Peer Groups. The procedure will search for the best number of peer groups between the specified minimum and maximum values. The values must be positive integers, and the minimum must not exceed the maximum. When the specified values are equal, the procedure assumes a fixed number of peer groups.

Note: Depending on the amount of variation in your data, there may be situations in which the number of peer groups that the data can support is less than the number specified as the minimum. In such a situation, the procedure may produce a smaller number of peer groups.

Maximum Number of Reasons. A reason consists of the variable impact measure, the variable name for this reason, the value of the variable, and the value of the corresponding peer group. Specify a non-negative integer; if this value equals or exceeds the number of processed variables that are used in the analysis, all variables are shown.

DETECTANOMALY Command Additional Features

The command syntax language also allows you to:

- Omit a few variables in the active dataset from analysis without explicitly specifying all of the analysis variables (using the `EXCEPT` subcommand).

- Specify an adjustment to balance the influence of continuous and categorical variables (using the `MLWEIGHT` keyword on the `CRITERIA` subcommand).

See the *Command Syntax Reference* for complete syntax information.

Chapter 6. Optimal Binning

The Optimal Binning procedure discretizes one or more scale variables (referred to henceforth as **binning input variables**) by distributing the values of each variable into bins. Bin formation is optimal with respect to a categorical guide variable that "supervises" the binning process. Bins can then be used instead of the original data values for further analysis.

Examples. Reducing the number of distinct values a variable takes has a number of uses, including:

- Data requirements of other procedures. Discretized variables can be treated as categorical for use in procedures that require categorical variables. For example, the Crosstabs procedure requires that all variables be categorical.
- Data privacy. Reporting binned values instead of actual values can help safeguard the privacy of your data sources. The Optimal Binning procedure can guide the choice of bins.
- Speed performance. Some procedures are more efficient when working with a reduced number of distinct values. For example, the speed of Multinomial Logistic Regression can be improved using discretized variables.
- Uncovering complete or quasi-complete separation of data.

Optimal versus Visual Binning. The Visual Binning dialog boxes offer several automatic methods for creating bins without the use of a guide variable. These "unsupervised" rules are useful for producing descriptive statistics, such as frequency tables, but Optimal Binning is superior when your end goal is to produce a predictive model.

Output. The procedure produces tables of cutpoints for the bins and descriptive statistics for each binning input variable. Additionally, you can save new variables to the active dataset containing the binned values of the binning input variables and save the binning rules as command syntax for use in discretizing new data.

Optimal Binning Data Considerations

Data. This procedure expects the binning input variables to be scale, numeric variables. The guide variable should be categorical and can be string or numeric.

To Obtain Optimal Binning

1. From the menus choose:

   Transform > **Optimal Binning...**
2. Select one or more binning input variables.
3. Select a guide variable.

Variables containing the binned data values are not generated by default. Use the Save tab to save these variables.

Optimal Binning Output

The Output tab controls the display of the results.

- **Endpoints for bins.** Displays the set of endpoints for each binning input variable.
- **Descriptive statistics for variables that are binned.** For each binning input variable, this option displays the number of cases with valid values, the number of cases with missing values, the number of distinct valid values, and the minimum and maximum values. For the guide variable, this option displays the class distribution for each related binning input variable.

© Copyright IBM Corporation 1989, 2016

- **Model entropy for variables that are binned.** For each binning input variable, this option displays a measure of the predictive accuracy of the variable with respect to the guide variable.

Optimal Binning Save

Save Variables to Active Dataset. Variables containing the binned data values can be used in place of the original variables in further analysis.

Save Binning Rules as Syntax. Generates command syntax that can be used to bin other datasets. The recoding rules are based on the cutpoints determined by the binning algorithm.

Optimal Binning Missing Values

The Missing Values tab specifies whether missing values are handled using listwise or pairwise deletion. User-missing values are always treated as invalid. When recoding the original variable values into a new variable, user-missing values are converted to system-missing.

- **Pairwise.** This option operates on each guide and binning input variable pair. The procedure will make use of all cases with nonmissing values on the guide and binning input variable.
- **Listwise** This option operates across all variables specified on the Variables tab. If any variable is missing for a case, the entire case is excluded.

Optimal Binning Options

Preprocessing. "Pre-binning" binning input variables with many distinct values can improve processing time without a great sacrifice in the quality of the final bins. The maximum number of bins gives an upper bound on the number of bins created. Thus, if you specify 1000 as the maximum but a binning input variable has less than 1000 distinct values, the number of preprocessed bins created for the binning input variable will equal the number of distinct values in the binning input variable.

Sparsely Populated Bins. Occasionally, the procedure may produce bins with very few cases. The following strategy deletes these pseudo cutpoints:

For a given variable, suppose that the algorithm found n_{final} cutpoints and thus $n_{final}+1$ bins. For bins i = 2, ..., n_{final} (the second lowest-valued bin through the second highest-valued bin), compute

$$\frac{sizeof(b_i)}{\min(sizeof(b_{i-1}), sizeof(b_{i+1}))}$$

where *sizeof(b)* is the number of cases in the bin.

When this value is less than the specified merging threshold, b_i is considered sparsely populated and is merged with b_{i-1} or b_{i+1}, whichever has the lower class information entropy.

The procedure makes a single pass through the bins.

Bin Endpoints. This option specifies how the lower limit of an interval is defined. Since the procedure automatically determines the values of the cutpoints, this is largely a matter of preference.

First (Lowest) / Last (Highest) Bin. These options specify how the minimum and maximum cutpoints for each binning input variable are defined. Generally, the procedure assumes that the binning input variables can take any value on the real number line, but if you have some theoretical or practical reason for limiting the range, you can bound it by the lowest / highest values.

OPTIMAL BINNING Command Additional Features

The command syntax language also allows you to:

- Perform unsupervised binning via the equal frequencies method (using the `CRITERIA` subcommand).

See the *Command Syntax Reference* for complete syntax information.

Notices

This information was developed for products and services offered in the US. This material might be available from IBM in other languages. However, you may be required to own a copy of the product or product version in that language in order to access it.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

For license inquiries regarding double-byte (DBCS) information, contact the IBM Intellectual Property Department in your country or send inquiries, in writing, to:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, Japan

INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you provide in any way it believes appropriate without incurring any obligation to you.

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The licensed program described in this document and all licensed material available for it are provided by IBM under terms of the IBM Customer Agreement, IBM International Program License Agreement or any equivalent agreement between us.

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBMproducts. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to actual people or business enterprises is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

Each copy or any portion of these sample programs or any derivative work, must include a copyright notice as follows:

© your company name) (year). Portions of this code are derived from IBM Corp. Sample Programs.

© Copyright IBM Corp. _enter the year or years_. All rights reserved.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at "Copyright and trademark information" at www.ibm.com/legal/copytrade.shtml.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

Index

Printed in USA