# Table of Contents

Getting Started ........................................................................................................ 2  
Getting Started with Analytics .................................................................................. 2  
System Requirements ............................................................................................... 3  
WORKSPACE Tab ..................................................................................................... 4  
Workspace Tab.......................................................................................................... 4  
Toolbar Menu ........................................................................................................... 4  
Open a File ............................................................................................................... 7  
Recent Section .......................................................................................................... 7  
Recent Analyses ....................................................................................................... 7  
Recent Connections ................................................................................................. 8  
Data Connections ..................................................................................................... 8  
Data Connections..................................................................................................... 8  
Data Connections Source Types .............................................................................. 8  
New Connection ....................................................................................................... 9  
Edit Connection ....................................................................................................... 9  
Delete Connection ................................................................................................. 10  
Selecting Tables for a Data Connection .................................................................. 10  
Defining Text Data Sources .................................................................................... 11  
Load Data From Source .......................................................................................... 14  
Load Data Options .................................................................................................. 14  
Opening Analyses from the Windows Command Line ............................................ 16  
Analyze Info ............................................................................................................. 17  
Analyze Info View ................................................................................................... 17  
Transform Variables ............................................................................................... 18  
Binning ...................................................................................................................... 18  
Multi-Variable Formula ......................................................................................... 19  
IF Function .............................................................................................................. 20  
Functions Operations ............................................................................................. 20  
Filter Records ......................................................................................................... 22  
Defining a Test .......................................................................................................... 22  
Filter Operations ..................................................................................................... 23  
STATISTICS Screen ................................................................................................ 24  
Statistics Screen ...................................................................................................... 24  
VIEW DATA Screen ................................................................................................ 26  
View Data Screen .................................................................................................... 26  
VISUALIZE Screen .................................................................................................. 27  
Visualize Screen ..................................................................................................... 27  
Univariate / Multivariate Charts ............................................................................ 28  
Univariate - Multivariate Options .......................................................................... 28  
Variable (X Axis) ..................................................................................................... 29  
Y Axis ......................................................................................................................... 29  
View By ..................................................................................................................... 29  
Show Related Variables ......................................................................................... 29  
Chart Types .............................................................................................................. 29  
Visualization Options Sidebar ................................................................................ 30  
Statistical Tests Option ......................................................................................... 31  

iii
Model Visualization Frame .................................................. 59
Decile Lift ............................................................................. 59
Decile Analysis ....................................................................... 60
Cumulative Lift ....................................................................... 60
ROC Curve Tab ........................................................................ 61
Report Screen ........................................................................ 63
Add to Report ........................................................................ 63
Saving the Model ..................................................................... 64
Memorize Model ...................................................................... 64
What If ..................................................................................... 64
Create Model Variable ............................................................. 65
Save Models ............................................................................ 65
Report Screen ........................................................................ 66
Add to Report ........................................................................ 66
Report Screen ........................................................................ 66
SCORING .................................................................................. 67
Score a File .............................................................................. 67
OPEN FILE ................................................................................ 67
SCORE TAB .............................................................................. 68
LOAD SCORING MODEL .......................................................... 68
SELECT MEMORIZED MODEL ................................................ 69
OUTPUT OPTIONS .................................................................... 69
START SCORING ..................................................................... 71
APPENDIX A - TASK FUNCTIONS SUMMARY ............................... 72
Function Descriptions (Multi-Variable Transform Task) ............. 72
Function Operators ................................................................... 76
Date and Time Format Specifiers ............................................ 77
Special Parameters ................................................................... 78
Contact Us ............................................................................... 78
Glossary of Terms ................................................................... 79
Definition Akaike Information Criterion (AIC) ......................... 79
Definition Autoline .................................................................... 79
Definition Binary Variable ....................................................... 79
Definition Categorical Variable ................................................. 80
Definition Concordance ........................................................... 80
Definition Continuous Variable ................................................. 80
Definition Correlation Analysis ................................................ 80
Definition Cumulative Lift ........................................................ 80
Definition Decile Analysis ........................................................ 80
Definition Delimiter .................................................................. 81
Definition Delta-P ..................................................................... 81
Definition Discordance ............................................................. 81
Definition G-K Gamma .............................................................. 81
Definition Hold Out Sample ..................................................... 81
Definition Intercept ................................................................... 82
Definition Kendall Tau A .......................................................... 82
Definition Log Likelihood Function ............................................ 82
Definition Logistic Regression ................................................... 82
Definition Logs Ratio Estimate .................................................. 82
Definition Mean ........................................................................ 82
Getting Started

Getting Started with Analytics

Rapid Insight Analytics™ is a software tool for analyzing analytic data sets and building predictive models.

Predictive modeling is an analytic approach that identifies patterns and relationships between data variables. These relationships are statistically validated and then used to build analytic predictive models. Such models are used to evaluate new data and apply a probability scores to each record, predicting the likelihood of a specific event or condition to occur. This product can be used alone or in conjunction with Veera™ (our data processing and reporting tool).

The Analytics Workstation is the core of the program. This is where all aspects of data analysis and predictive model building are accomplished. The Analytics Workstation is comprised of nine elements:

**Workspace Tab**
Where data sources and prior analyses are managed. This is also where new analyses are begun.

**Statistics Tab**
Allows users to view simple statistics for each of the variables, giving a clear and concise way of gaining insight into the dataset.

**View Data Tab**
Provides the user with a view of the entire dataset within the software.

**Visualize Tab**
Data are graphically rendered and optionally incorporated into reports

**Correlation Tab**
Displays positive and negative correlations between variables

**Analyze Tab**
A variety of different analyses may be performed on the selected dataset from this tab

**Clustering Tab**
The clustering feature is a way to categorize records (bins) in your dataset, using one or more variables as the criteria for the binning process.

**Model Tab**
Analytics identifies those factors related to a chosen variable (‘Y’), then produces one or more predictive, statistical models.

**Report Tab**
Where you are able to add and customize visualizations for an exportable report.
System Requirements

Systems intending to host Analytics must meet the following hardware requirements:

- Windows 7, SP1 or later
- .NET Framework 4.6.1 and its requirements
- 2 GHz 32-bit (x86) or 64-bit (x64) processor
- 1 GB of system memory
- 4.5GB available hard drive space
- 32 MB of graphics memory
- Internet access
WORKSPACE Tab

Workspace Tab

The Workspace tab automatically opens when the program first starts. It serves as the place where users manage data source connections and prior data analyses as well as where new data analyses are begun.

Toolbar Menu

The toolbar menu appearing at the top of the window provides options affecting the overall product.
FILE

- **Open Analysis**

Users selecting this option will identify a saved Rapid Insight analysis file (*.ria) and open it along with the dataset used to originally develop it. The dataset file/table must be in its original location, must be named the same and must contain the same variables. The program will automatically open to the screen where the original analysis was saved.

- **Save Analysis / Save Analysis As**

To save all parts of the current analysis, click “Save Analysis” under the “File” menu at any time.

- **Exit**

Exit closes the program. You will be prompted to save any unsaved analyses at this time.

TOOLS (Options)

- **General**

  Provides user access to several product-wide user interface settings.

- **Visualization**

  Defines the colors to be used in visualizations (charts & graphs)

- **Correlation**
Opt to display or not display P-values within the correlation tab.

- **Modeling**
  Configures standard modeling defaults such as default P-values.

- **Decision Trees**
  Provides settings for decision tree customization such as segment size.

- **Automine**
  Configures standard "automine" conditions and criteria such as category handling.

- **Reporting**
  Opt to add or remove the "powered by Rapid Insight" report watermark.

**VIEW (Options)**

- **Reset Layout**
  Resets all customized options to their original values.

**HELP**

- **Contents**
  Provides access to the built-in product help.

- **Log Files**
Analytics logs any errors it encounters, along with troubleshooting details, into daily text files. This menu item provides user access to these files.

- **License and Register**

  Product licensing is accomplished here.

- **Updates**

  Users may check the web for the latest product updates along with details about prior updates.

- **About**

  Product details are stored here.

---

**Open a File**

Clicking this button will allow the user to use Windows Explorer to locate and select a data file for analysis.

---

**Recent Section**

**Recent Analyses**

The last 10 analyses saved by the user will appear in this frame. Double-clicking on an entry will open the corresponding Rapid Insight analysis file (*.riax) along with the dataset used to originally develop it. The dataset file/table must be in its original location, must be named the same and must contain the same variables. The program will automatically open to the screen where the original analysis was saved.
**Recent Connections**

The Recent Connections frame lists the last ten unique data sources accessed by the program. When an entry in the list is selected, two icons appear on either side of the listing.

- The pushpin (on the left) fixes that entry in place. It will not be displaced by other entries until "unpinned".
- Clicking the "Load" button (on the right), or double-clicking the entry, will begin the process of loading the data into the program and starting a new analysis (see [Load Data From Source](#)). Note that the dataset file/table must be in its original location, must be named the same and must contain the same variables. The program will automatically open to the screen where the original analysis was saved.

**Data Connections**

**Data Connections**

A data connection (or just “connection”) can be thought of as a pointer. It shows Analytics where data is located on the local computer, another computer on the local network, or a cloud-based resource. It can reference either a file folder or a database on a server. Once created, a connection may be used over and over again by any number of different jobs. The Connections section of the Workspace tab is where connections used by the product appear and can be managed (created, modified or deleted).

**Data Connections Source Types**

Connections are automatically grouped by data source type. Data source types Analytics can support include:

- Access
- DB2
- DSN (ODBC Connections)
- Excel
- FoxPro
- Hadoop
- MySQL
- Oracle
Which data source types are available is often dependent upon whether the drivers or providers for that type are installed on the computer. Drivers (OLE DB, ODBC, ADO.net or native drivers) are freely available for download from the corresponding manufacturer.

New Connection

Creating a Data Connection

Note that creating a data connection does not actually create any folders, files or tables. It only creates a reference within Analytics pointing to an existing folder on the host computer or network.

- Open the Create New Connection dialog box by doing one of the following:
  - Right-click in any Jobs > Connections section and select ‘New Connection’, or
  - Clicking on the ‘New Connection’ link.
- Enter a unique Connection Name.
- Select a Data Source by either
  - Clicking the icon of the corresponding data type, or
  - Selecting the corresponding data type from the drop down list.
- The dialog box changes to request the information unique to each type. This means either
  - Identifying the folder location, or
  - Identifying the database server location, database name and connection information.

Edit Connection
There are two ways to edit a specific data connection.

- Select the data connection from the list, then click the Edit Connection link at the top of the Connections frame, or
- Right-click on the data connection and select "Edit Connection" from the menu that appears.

The connection properties will appear and be open for editing.

Delete Connection

There are two ways to delete a specific data connection.

- Select the data connection from the list, then click the Delete Connection link at the top of the Connections frame, or
- Right-click on the data connection and select "Delete Connection" from the menu that appears.

In both cases, a confirmation window will appear; click YES to complete the deletion. Once a connection is deleted, this action cannot be reversed.

Selecting Tables for a Data Connection

Once a data connection is saved, the user can create references to the specific files, worksheets, or database tables/views found in that location. These specific references are collectively referred to as “tables”. Tables appear beneath their parent connections in alphabetic order. To add a table:

- Select the data connection, then either
  - Click the Select Tables link at the top of the Connections frame, or
  - Right-click on the data selection and choose select tables.
- Select the file(s), worksheets, etc. to be used from the connection location.
For Text file data connections, the properties often need to be defined (see Defining Text Data Sources).

If the desired table does not appear in the Files/Table list, right-click in the space and click Add Table By Name to manually enter the file/table name into the list (not available for Excel or Text).

Defining Text Data Sources

If a text type data connection is saved, the user will then need to create references to the specific text files that will serve as data sources. If the text file selected does not have an [X] in it’s ‘Has Properties’ checkbox, the properties need to be defined. (NOTE – typically, the defaults that appear on the following three windows will be acceptable and no changes will need to be made.) Click the Properties button associated with the file to start the process.

File Properties Tab (within text connection – select tables window)

The first tab of the File Properties window defines the file’s general data formatting. These include:

- **ROW DELIMITER**: Determines what constitutes a new data row. Options are: {CR} (carriage return), {LF} (line feed), {CR}{LF} (default), comma, semicolon, tab, and vertical bar.
- **FILE TYPE**: Special format types used under special circumstances. Options are: Default, OEM, ASCII, Unicode, UTF8, UTF7
- **TEXT QUALIFIER**: Determines how data that contains a column delimiter is quoted. Options are: Double Quote (default), Single quote, and none.
- **SAMPLE SIZE**: Defines the number of data rows to load to evaluate the structure. Default is 20.
- **DELIMITED or FIXED WIDTH**: Determines how the data will be separated into columns.

### Delimited Fields Tab

![Delimited Fields Tab](image)

This tab appears only if the DELIMITED option was chosen on the File Properties Tab. Provides a sample of the data set using the Field Delimiter chosen.

- **FIELD DELIMITER**: Options include: comma (default), semicolon, tab, vertical bar, tab, colon, and other – with entry field.
- **FIRST ROW CONTAINS FIELD NAMES**: Yes or no checkbox.
- **COLUMN COUNT**: Drops existing or adds new data columns. Default is based on the current column count detected by the delimiter selected.
- **IMPORT FORMAT FILE**: If the user chooses to set up the data columns automatically, the user needs to import a format file (.fmt) with the desired parameters. To start the process, the user clicks on the IMPORT button.

### Fixed Width Fields Tab

This tab appears only if the FIXED WIDTH option was chosen on the File Properties Tab. The user may manually set the data column’s start and end points or may import a format schema to do this automatically.
- **FIELD POSITIONS – RECORD WIDTH**: sets the intended width in characters of each row of data.
- **FIELD POSITIONS – FIELD**: To manually set column start and end points, the user clicks on the point in the sample data where a data column starts. Columns will alternately shade as they are added. User may remove a column pointer by clicking on the same point. The TOTAL COLUMNS counter keeps track of the columns designated.
- **IMPORT FORMAT FILE**: If the user chooses to set up the data columns automatically, the user needs to import a format file (.fmt) with the desired parameters. To start the process, the user clicks on the IMPORT button.

**Column Properties Tab**

This final tab of the File Properties window defines the data types and labels of each data column. The user clicks on the column from the sample data field and then sets the following values.
• CURRENT COLUMN: Alternate method for selecting the column being configured.
• NAME: Column’s name.
• DATA TYPE: The supported types include Integer, Text, Date, Yes/No, Unknown, and Decimal.
• DATETIME FORMAT: If the column’s data type is DATE, this field sets the date formats as they will appear in Veera coming from this source.
• DECIMAL PLACES: If the column’s data type is DECIMAL, this field sets the number of digits as they will appear in Veera coming from this source.
• INSERT/DELETE COLUMN: Drops existing or adds new data columns which can then be configured.

Load Data From Source

Once a data source has been selected for analysis, this dialog window opens:

Clicking the "Options" arrow button will reveal the variety of options that may be imposed on the dataset as it is loaded for analysis. Please see the relevant help sections for a description of each option.

Clicking the "Load Data" button will apply the selected options and load the records before beginning the analysis.

Load Data Options

A variety of options that may be imposed on the dataset as it is loaded for analysis.

% of Data to Load

This control is only active prior to clicking the Load Variables button. It allows the user the option of analyzing just a subset (1
15%

Randomly Order Rows

If the Random Order Rows box is checked and the % of Data to Load is set to something other than 100%, the records to analyze will be selected randomly. Otherwise, they will be selected in the order encountered in the file/table.

Max Category Count

A user-defined value capping the number of unique categories a variable may contain and still be considered a categorical variable. If this number is exceeded, the variable is reclassified as a text variable (if one or more of the values contain letters) or a numeric variable (if all the values consist exclusively of numbers).

Category Min Occurrence

A category must appear this user-defined number of times in the dataset to be considered worthy of individual analysis.

Exclude Columns & Identifier Columns

Variables to Exclude from Loading

This field is only active prior to clicking the Load Variables button. Variables that the user has determined should not be part of the analysis are selected from the list on the left and "moved" to this field using the Exclude button. Variables may be returned to the list of loadable columns by selecting them and clicking the Include button.
Mark Variables as Identifiers

This field is only active prior to clicking the Load Variables button. Variables that serve as unique record identifiers in the dataset should not be part of the analysis. To keep them from loading, the user selects them from the list on the left and “moves” them to this field using the Identifier checkboxes. Variables may be returned to the list of loadable columns by selecting them and clicking the checkboxes again.

Load Data

Once clicked, the program will automatically open an Analysis tab for the chosen dataset. Any options selected will be applied to the records before the analysis begins.

Opening Analyses from the Windows Command Line

Analytics analyses may be opened automatically using a command executed from the Command Line Prompt or the RUN function under the Windows Start menu. The command itself is comprised of two parts. The first is the file path where the Analytics program file "AnalyticsWorkstation.exe" is located. On a machine using the Windows 7 OS, this would be:

\textit{C:\Program Files (x86)\Rapid Insight Inc\Analytics\RapidInsight.exe}

Next would be the actual command to load the analysis ("\texttt{O }"), followed by the path and the specific name of the analysis (i.e. \texttt{C:\Temp\Enrollment Analysis.riax}) A fully-assembled command would read something like this...

\texttt{C:\Program Files (x86)\Rapid Insight Inc\Analytics\RapidInsight.exe O C:\Temp\Enrollment Analysis.riax}
**Analyze Info**

**Analyze Info View**

Users can view general characteristics of the dataset and its analysis by clicking this downward-pointing "expand" control. To hide these values, simply click the upward-pointing "collapse" control. These characteristics are available in all Workstation Tabs.

Dataset characteristics reported on include:

- Total Records in DataSet
- Records Using (after any filtering)
- Percent Loaded (optionally configured on the Load Data screen prior to the start of the analysis)
- Current Filter (denotes any filters being applied)
- Variables in Dataset (original count)
- Variables in Analysis (subtracting those columns optionally set aside on the Load Data screen prior to the start of the analysis and adding any variables developed by the user using Analytics' "Transform" function)
- DataSet Last Loaded On (date and time)
- Y Variable (with an opportunity to select either a binary or continuous variable)
Transform Variables

Overview

The Transform Node allows you to create new columns from existing columns. It will show a list of all of the fields that exist in the analysis. There are several options for different types of variables that can be created. Along the top of the Transform Window, there are four tabs (Binning, Multi-Variable Formula, Text Functions, Date/Time).

Binning

The binning option allows users to create bins from another variable. To do this, first select the column that users want to create the bins with, on the left side. Next name the first bin/category. To do that, simply type in a name for that bin where it says ‘First Category =’. Next, define the definition for that bin.

For example, perhaps users have a field containing SAT Math scores, and want to define ‘low’, ‘medium’, and ‘high’ bins. In that case, name the first category ‘low’. Then define that category as (SAT Math <= 500).

Once defined, click the ‘validate’ button. Then name and define the next category.

Perhaps a user wants to name the next category as ‘medium’, and define it as SAT Math > 500 and SAT Math <=600. Click validate again, and define the next category. Once all categories are defined, type the name of the new variable where it says ‘New Variable Name’.

Finally, click the ‘create variable’ button. This new Transform now appears at the bottom of the window where it says ‘Transform Operations’. The Transform can be edited at any time by clicking it, making any changes, and then clicking the ‘update’ button.
A good way to test whether the transform was created correctly, without having to run the entire job, is to save and exit the Transform Node, then right click the Transform Node and choose ‘preview data’. By default, this runs the job using the first 1,000 records from any inputs, and shows the results, with the new variable as the last variable.

Note: If the variable being used to create the binned variable is a categorical variable like ‘state’, users can retrieve the unique values by clicking the ‘get sample’ button in the upper right corner. Users can then choose the ‘in’ operator, hold down the shift or control keys and choose any values to be placed into each bin.

Multi-Variable Formula

The contents of a new column can be derived from a single, solvable formula using one or more existing columns.

- First, choose any columns to be used as variables in the formula. Veera assigns formula tags to each of the chosen columns.
- In the formula box, any formula can be entered. In the example A+B is typed. This adds the two fields.
- Next choose a Result Type from the drop-down box.
- Finally, name the new variable, and click Create.
IF Function

The IF function returns one value if a condition specified evaluates to TRUE, and another value if that condition evaluates to FALSE. The syntax for the IF function is:

IF(P1,P2,P3), where:

- P1 is the condition to test
- P2 is the value to return if the expression is true
- P3 is the value to return if the expression is false

As an example IF([A]>50,'high','low') would return the value of 'high' if the field represented by A was greater than 50, and 'low' if it is not greater than 50. Note in this case, since the output values are text, single quotes are put on 'high' and 'low', and the result type would need to be set to text.

Functions Operations

Clicking the f(?) button displays a list of functions. Clicking a function within that list shows the definition and parameter requirements of the function. The function can either be manually entered into the formula box, or double clicked to automatically appear in the formula box.

MATCHESREGEX Function

MatchesRegEx is a function that takes a text column and returned a 1 when the text string matches the desired regular expression (RegEx). In effect, MatchesRegEx acts as a universal IS function, although users must develop the testing expression. Fortunately, searching the Internet for regex provides a wealth of examples. For example:

- This RegEx will only match a valid social security number (i.e. it won't match 000-00-0000): `^(?!000)(([0-6]\d{2}|[7([0-6]\d|7\d|7[012]))([ -]?)(?!00)\d\d\d(?!0000)\d\d\d)$`
- This RegEx will match currency (with or without cents, and optional $ sign): `^[\d+\.(\d{2})\d\d?\$`
- This RegEx will match positive integer: `^\d+$`

Boolean Operations

Users can create Boolean (true/false or 0/1) variables. Choose the columns to work with, to assign variable tags. Type an expression and choose the output type of either Boolean which results in a true or a false output, or choose Binary, which results in 0 or 1 output.
Validate Formula

This button appears above the formula field. It checks the contents of the formula field for possible syntax issues. If a potential problem is detected, an Invalid Formula error is displayed. No message appears if the syntax is successfully validated.

Text Functions

Only TEXT columns are available on the Text Functions tab. Choose the text column to work with then fill in any parameters that are requested (the function definition as well as the parameter definitions are contained in the Function Definition box). Once the required parameters are filled in, name the new transform and click the Create button. If no text columns are found, the column list is blank.

Date/Time Functions

Only DATE/TIME columns are available on the Date/Time tab. Choose the date field to work with from the left. Next, fill in any parameters that are requested (the function definition as well as the parameter definitions are contained in the Function Definition box). Once the required parameters are filled in, name the new transform then click the Create button. If no date/time columns are found, the column list is blank.

Editing a Transform Operation

Transforms can be edited at any time by clicking the transform at the bottom of the window. Once the edits are made, click the Update button to update the transform. A transform can also be removed by clicking the transform and clicking the Delete button.

Copying a Transform Operation

Formulas and bins may be copied by right-clicking the entry in the Transform Operation list at the bottom of the window and selecting Clone Transform. An identical copy of that operation will appear in the list.

Exporting and Importing a Transform Operation

Formulas and bins may be exported by right-clicking the entry in the Transform Operation list at the bottom of the window and selecting Export Transform. The program will allow the operation to be saved as a *.ritd file.

Formulas and bins may also be imported by right-clicking in the Transform Operation list at the bottom of the window and selecting Import Transform. The program will allow the user to select a saved *.ritd file and add its operation to the existing list.
Filter Records

Overview

Filter is used to filter records based on one or more user-defined tests.

Defining a Test

Clicking on a column name from the list on the left selects it as the source of the data to test. In the example shown, "State" has been selected. Then the type of test is chosen from the following options:

- In (specified list)
- Not In (specified list)
- Contains (found within column value)
- Does Not Contain (not found within column value)
- Starts With (the specified value)
- Does Not Start With (the specified value)
- Ends With (the specified value)
- Does Not End With (the specified value)
- Like (matches a pattern)
- Equals (=)
- Not Equals (<>)
- Greater Than (>)
- Less Than (<)
- Greater Than or Equal To (>=)
- Less Than or Equal To (<=)
- Between (two specified values)
- Is Missing (no value exists)
- Is Not Missing (a value exists)
- EXPRESSION (offers range of DATE-specific operators that rely upon the computer system clock for reference, such as TODAY). This test is only available when the column selected is a 'date/time' type variable.

Next, the user must enter the data value(s) to test for. As an aid, a “Get Values” button is provided, permitting the user to view unique values found in the column being tested. Values may be selected directly from this list (holding the Ctrl key allows multiple selections).

Filter Operations

Creating Multiple Tests

Once a test is defined, it is added to the filter by clicking the Add Filter button (green funnel). It will then appear in the area just below the control. A corresponding SQL statement describing the test will also appear in the bottom-most box. An existing test may be removed from this area by highlighting it, then clicking the Delete Filter button (blue funnel).

Multiple tests may be combined to construct sophisticated filtering criteria. The following tools are provided to facilitate this.

Insert/Remove Parenthesis

Parentheses may be useful when combining several tests. The parentheses are used to group test criteria together so they can run together instead of separately. The parentheses can also be removed. Select each one, then use the arrow next to each parenthesis and choose the remove option.

And/Or Statements

Multiple tests may be joined together by adding AND or OR between them. To change one to the other, select the value then click the desired button to change it. Selecting the OR Otherwise, by default, tests are initially joined by AND

Move Up/Down

If there is more than one test, their order can be changed using this tool. Select the test and click the move up or move down arrow.

Delete All Filters

This deletes all of the test criteria from the filter.
STATISTICS Screen

Statistics Screen

The Statistics tab allows users to view simple statistics for each of the variables, giving a clear and concise way of gaining insight into the dataset. It is a good idea to check these statistics because problems such as data omissions and outlying points are often detected here.

This grid contains information about each variable’s mean, standard deviation, minimum value, maximum value, data type, coefficient of variation, number of missing values, number of observations, number of distinct values, range, and whether a variable is text or not.

- **Type (variable type)**
  
  This represents the type of variable (continuous, binary, text, etc.).

- **Mean**
  
  The mean is the average value of all of the observations of one numeric variable.

- **Std Dev (standard deviation)**
  
  The standard deviation represents the “spread” of data around its mean. A higher value indicates a more spread out data set than a lower value.
• **Min (minimum)**
  The minimum value represents the smallest observation found.

• **Max (maximum)**
  The maximum value represents the largest observation found.

• **Coeff Var (coefficient variation)**
  This represents the ratio of the standard deviation to the mean. The larger this number is, the larger the standard deviation is relative to the mean.

• **Missing**
  This number denotes the number of missing (null) observations for each variable.

• **# Obs (total # of observations)**
  This is the total number of non-null observations for each variable.

• **Range**
  The range represents the difference between the highest and lowest observation for each variable.

• **# Distinct (categories)**
  This represents the number of unique categories within each variable.
VIEW DATA Screen

View Data Screen

Use the View Data tab to review the entire dataset.

Edit Sort

The "Edit Sort" (A-Z button) opens a pop-up box that offers the user a selection of sorting and search options.

- A-Z button sorts the data in ascending order.
- The Z-A option sorts the data in descending order.
- The AxZ option unsorts the data, putting in its original order.
- The binoculars icon bring up a dialog box to search for a particular column if there are a number of entries.

Reset Sort Order

This button removes all prior sort settings and returns the display to its default (actual) order.
VISUALIZE Screen

As the name suggests, this corner of the Analytics program is dedicated to visually representing the data records under analysis. There are two fundamental viewing modes - univariate and multivariate. These, in turn, determine what presentation options are made available to the user. Note that some additional display choices (such as what colors to use) are available under the TOOLS > OPTIONS menu at the top of the program.

Choose Y and X Variables
Univariate / Multivariate Charts

Univariate - Multivariate Options

Selecting the Univariate Analysis visualization type option will permit a chart representing a single variable to be displayed. View the distribution of any variable in the dataset by simply selecting the variable from the list.

Selecting the Multivariate Analysis visualization type option allows us to analyze two or more variables simultaneously, illustrating how pairs of variables may be related.
Variable (X Axis)

To generate a chart, select a variable from the drop down list, either titled 'Variable for Univariate, or 'Related Variables' for Multivariate. The chosen variable will be positioned on the X-axis.

If the visualization type is univariate, the chart will display immediately. If the visualization type is multivariate, the chart will not be displayed until the Y-axis variable is selected.

Y Axis

This field appears only when the multivariate visualization option is chosen. The variable selected from the drop down list will be placed on the Y-axis and its relationship to the X-axis variables) graphically depicted.

View By

This field appears only when the multivariate visualization option is chosen and both the X-Axis and Y-Axis fields are populated. The variable selected from the drop down list will be placed on the X-axis as a sub-category and its relationship to both the X-Axis and Y-Axis variables graphically depicted.

Show Related Variables

This control allows users to establish a p-value threshold for variables that will be visually represented.

Chart Types

Analytics offers three chart types to graphically represent data - pie, bar and bubble.

BAR CHARTS
Bar charts visually represent data as vertical bars against a numeric scale displayed on the Y-Axis.

PIE CHARTS
Pie charts visually represent data as relative, proportional slices of a circle.

BUBBLE CHARTS
Bubble charts visually represent data elements as proportionally-sized and relatively-tinted circles.
Visualization Options Sidebar

This sidebar tab is available in both univariate and multivariate viewing modes. It provides an array of self-described options that affect display elements such as color, fonts, labeling, and data order. The options themselves change depending on which viewing mode the user has selected.

This view may be "pinned" permanently open by clicking the pushpin icon in the upper right corner of the sidebar frame. Clicking the pushpin again will unpin the sidebar so it will retract when not in use.
Statistical Tests Option

This statistical option is available only when in the multivariate viewing mode, by selecting the ‘Statistical Tests’ checkbox within the Visualization Options Sidebar. It provides additional detail describing the statistical relationships between the data underlying the chart in view.

<table>
<thead>
<tr>
<th>Category</th>
<th>Related</th>
<th>F Value</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>STAT</td>
<td>True</td>
<td>6.709</td>
<td>0.041</td>
</tr>
<tr>
<td>POLT</td>
<td>True</td>
<td>5.299</td>
<td>-0.036</td>
</tr>
<tr>
<td>CMN</td>
<td>True</td>
<td>5.111</td>
<td>-0.036</td>
</tr>
<tr>
<td>ECON</td>
<td>True</td>
<td>4.812</td>
<td>-0.035</td>
</tr>
<tr>
<td>RUSS</td>
<td>False</td>
<td>11.089</td>
<td>0.053</td>
</tr>
<tr>
<td>HT</td>
<td>False</td>
<td>6.983</td>
<td>0.042</td>
</tr>
<tr>
<td>GEOG</td>
<td>False</td>
<td>6.918</td>
<td>0.042</td>
</tr>
<tr>
<td>MICR</td>
<td>False</td>
<td>6.069</td>
<td>0.039</td>
</tr>
<tr>
<td>CLAS</td>
<td>False</td>
<td>5.813</td>
<td>-0.038</td>
</tr>
<tr>
<td>MLS</td>
<td>False</td>
<td>5.470</td>
<td>0.037</td>
</tr>
<tr>
<td>FREN</td>
<td>False</td>
<td>5.381</td>
<td>0.037</td>
</tr>
<tr>
<td>HMP</td>
<td>False</td>
<td>4.717</td>
<td>0.034</td>
</tr>
<tr>
<td>Pre-Md</td>
<td>False</td>
<td>3.611</td>
<td>0.030</td>
</tr>
<tr>
<td>NURS</td>
<td>False</td>
<td>3.385</td>
<td>0.029</td>
</tr>
<tr>
<td>WILD</td>
<td>False</td>
<td>2.708</td>
<td>0.026</td>
</tr>
</tbody>
</table>
Report Screen

Add to Report

The 'Add To Report' button appends the chart or graph in view to Analytics' built-in reporting feature. This is available within the Visualize Tab, the Clustering Tab and the Model Tab. To view the report itself, click on the "Report" Tab.

Report Screen

The Report Screen provides a handy place to incorporate any charts into one of several available reporting formats. When all of the desired charts have been assembled, clicking on the "Generate Report" ( ) control will publish the contents as a web page (.html), a slide show (.pptx), a spreadsheet (.xlsx) or a document (.docx or .pdf). The contents will remain in the Report Screen as long as the analysis is active or until removed using one of the delete slide controls ( ) at the top of the frame.

Individual graphics may individually viewed, annotated and saved (as a .png) by highlighting the desired graphic and clicking the "Open Selected Slide" ( ) control at the top of the frame.
Correlations Screen

The Correlation Analysis tab displays positive and negative correlations between variables. The Correlation Matrix displays the correlation coefficients for each variable pair. The correlations range from -1 to +1 and indicate the degree of association between two variables. A positive value implies a positive association while a negative value implies an inverse association. The closer each value is to -1 or +1, the stronger the degree of association.

Select All / Clear All / Select Variables

Use several of the controls found in the Columns List to adjust the number of variables displayed in the cross reference. Use Select All to see all of the variables and their statistics at once, Clear All to clear all variables from the table, or individually Select Variables by clicking the check boxes next to each one.
The Analyze view provides a series of tabs, each of which represents a unique type of data analysis. These include:

- Means
- Frequency
- Profiling

It also contains statistics pertaining to the dataset under analysis ("Analysis Info") as well as opportunities to create new variables from existing content ("Transform") and filter the dataset to focus on a subset of the broader population. Refer to the relevant help section below for details regarding these analyses and functions.
The left sidebar allows you to see summary statistics only for the selected variables. The statistics table adds or removes variables from view, making it easy to place variables next to each other to see how they compare.

### Show / Hide All

Show All displays all of the variables and their statistics at once; Hide All clears all variables from the table, while Select Variables individually displays variables by clicking the check box next to each one.

### Sorting Options

The A-Z option sorts the data in ascending order. The Z-A option sorts the data in descending order. The AxZ option un-sorts the data, putting it in its original order.

### Find

The binoculars bring up a dialog box to search for a particular column if there are a number of entries.
Means Tab

GENERAL DESCRIPTION

The Means Analysis tab performs means analysis on any continuous variable, which includes a mean, count of the observations, maximum value, and minimum value for each category or subclass. The means for each category of a categorical or binary variable can also be further sub-classed by a second categorical or binary variable.

Variables (numeric)

From the variable list, select a variable to examine the mean of by highlighting the variable and clicking the arrow to move it to the “Variables to examine means of” window. You may select more than one variable to examine the means of in this step. The means that appear in the output table are the means (categorized by any other variables selected) for the variables selected in this field.

By (optional)

Select one or more variables from the “Variables (binary and categorical)” list by highlighting them then clicking the arrow to move it to the “By (optional)” field. Click “Analyze” to view the results of the means analysis.
Frequency Tab

GENERAL DESCRIPTION

The Frequency Analysis tab looks at the frequency of occurrence for binary and categorical variables.

SINGLE VARIABLE CATEGORY FREQUENCIES

Users select one or more variables from the list provided using the arrow buttons. By clicking the “Analyze” button, the counts and percentages for each variable selected are displayed.

Selecting a "By" category further breaks down these variable frequencies by a variable category. There are four views for each comparison of two variables -

- raw numbers,
- row percentages,
- column percentages, and
- total percentages
Profiling Tab

GENERAL DESCRIPTION

The Profiling Analysis tab compares two groups of a binary variable to determine how statistically different they are.

SELECT A BINARY VARIABLE TO PROFILE

The “Select a Binary Variable to Profile” window lists all of the binary variables in a dataset. After highlighting a variable, create names for the binary variable’s two values in the Profile Work Area.

P-VALUE

The P-Value drop-down menu gives the ability to change the p-value according to how statistically significant the relationship between the X-variables and Y-variable should be. The lower the p-value, the more significant the relation.

PROFILE

Clicking the “Profile” button generates a chart that compares the categories of the binary variable chosen to variables for which the two categories have statistically significantly different means or proportions. The upper portion of the chart compares each category to each of the continuous variables in a dataset by showing the means of each variable for both categories of the binary variable. The lower portion of the chart compares each category to each of the binary and categorical variables in the dataset for which the two categories have statistically significantly different means or proportions. By checking the “Show Z Score” button, a Z-score (measure of how many standard deviations an observation is above or below a mean) is included. The larger the Z-score, the further away each observation is from the mean.
The clustering feature is a way to categorize records (bins) in your dataset, using one or more variables as the criteria for the binning process (see "Binning" below). It utilizes the “Lloyd’s K-Means Clustering” algorithm. There are several methods for binning records based on their observed values.
Clustering

In this case, “Lloyd’s Algorithm” forms “k” categories, where k is a number that the user specifies. The algorithm then assigns each record in your data into one of those categories. The perk of creating these categories through K-means, instead of one of the other methods, is that the algorithm’s clusters will take into consideration the distribution exhibited in your data. A manually generated binning scheme might include a “50 to 54” age category, but the K-means clustering would not create that category if the dataset has no records in that age range. K-means clusters reflect the data and is not at all arbitrary. For strict comparisons across analysis, you would need to be sure to import identical clustering models.

Binning

This refers to the process of grouping records in your dataset according to a specified rule. Binning uses the raw values found within columns to identify which larger group each record belongs to. This could be hard-coded, as in the case of fixed values, like “0-5”, “6-10”, “11-15” (see Transform Variables), or dynamic, as in the case of the K-Means Clustering. The end result is an additional column which serves as an over-arching grouping variable.

Clustering Models

This is a decision rule created by the Clustering process. It allocates records to appropriate bins, based on their values in whichever columns were referenced while clustering.

Clustering Analysis Controls/Options

- K (number of clusters):
  
  This dropdown lets you choose the number of “clusters” or “groupings” you want to form.

- X Axis (projected):
  
  This allows you to choose which of your “Included Variables” will be displayed along the x axis (horizontal) of the “Clustering Visualization.”
- **Y Axis (projected):**
  
  This allows you to choose which of your “Included Variables” will be displayed along the y axis (vertical) of the “Clustering Visualization.”

- **Granularity:**
  
  The number indicated in this box indicates the number of data points that will be displayed on the “Clustering Visualization.” You can adjust this number manually, or with the slider.

- **Generate Clusters:**
  
  Clicking this tells the software to create the desired number of clusters with respect to any variables that you’ve included.

- **Save Variable:**
  
  This adds your created cluster to the dataset, and enables you to use it for analyses in the various other tabs of Analytics

- **Save Model:**
  
  This feature lets you export a “clustering model” (.ricm) so that you can translate your cluster variable into a Veera job.

- **Clustering Visualization:**
  
  This graph plots data points with respect to their characteristics (which are labeled on the x and y axes). The color coding serves to identify which cluster the data point belongs to, while the black squares represent the mean associated with each cluster.

- **Computed Means:**
  
  This tabular readout identifies the record count of each cluster (the “Population” column) as well as the mean(s) of the variable(s) used in the clustering process.
MODEL Screen

This screen is where models to predict the Y-variable are built. Rapid Insight Analytics can build a model automatically or users can manually create their own custom models.

Analytics uses regression analysis to develop its predictive models. Regression uses several techniques for modeling and analyzing data describing past behavior. More specifically, regression analysis helps explain how the typical value of the Y-variable changes when any one of the other, independent variables is changed, while the other independent variables are held fixed. This is known as a “step”. These series of stepwise results are then combined into a formulaic whole – the predictive model. Analytics offers two types of regression analysis – Logistic and OLS (Ordinary Least Squares).
Selecting a Y Variable

Select one variable to be the focus of the modeling effort - otherwise known as the “Y Variable”. Click the "select a Y variable" link near the top of the view and select one of the variables displayed in the pop-up list. Only binary and continuous variables may be selected.

Automine Tab

The Automated Mining tab lets us take advantage of Rapid Insight Analytics’ ability to quickly evaluate all variables against the Y-variable to find any statistical relationships that exist. In order to accomplish this, a Y-variable must first be chosen in the Control Panel tab. The Automine process automatically color-codes the variables according to whether they are statistically related to the selected Y-variable.

P-Value

This control allows users to establish a p-value threshold for variables that will be considered for inclusion in the model.
Set Missing Handling

Set Missing Handling identifies missing values within each variable. If missing values are found, several options are available from the menu:

- **Skip Missing Values** (Note - it is important to keep in mind that excluding multiple records can greatly reduce your sample size for analysis; do so with caution.)
- **Fill with Mean**
- **Fill with Zero**
- **Default** (defined under the Default Options toolbar menu)

Set Model Availability

A variable’s Model Availability column determines whether a column will or will not appear in any predictive model formula. Users may select an option from a given list:

**Exclude**

Excludes a variable from modeling, even if it is related to the Y-variable,

**Exclude All**

Excludes all variables from modeling, even if it is related to the Y-variable,

**Available**

Make the variable “Available” even if it is not related to the Y-variable, and

**Available if Related to Y**

Allow the software to include the variable only “If Related to Y”

If the variable is data types text, date, or constant, the column will be blank and won’t be available for use in a model.
Model Building Tab

Modeling Options Sidebar

The Modeling Options sidebar offer the user the ability to set the conditions under which a predictive model is developed. There are three options screens corresponding to the three primary regression types offered in the product - OLS, Logistic and Decision Tree (or CHAID).

OLS & Logistic Modeling Options

Modeling options control how a model is built and what is displayed in the outcome.
**Missing Values Handling**

The Missing Handling column identifies any missing values within each variable. If missing values are found, an option must be selected from the drop-down menu. “Default” (which is defined on the “Control Panel” tab), “Fill with Variable Mean”, or “Exclude Entire Record”. It is important to keep in mind that excluding multiple records can greatly reduce sample size for analysis; do so with caution.

**Fill with Variable Means**

Clicking the “Fill with Variable Means” button will fill in the missing values for each variable with the average value for that variable.

**Exclude Entire Record**

Clicking the “Exclude Entire Record” will not replace a missing value with any other value; rather, it will throw out that record for all variables so that it will not affect the analysis.

**P-Values for Stepwise**

The “P-Values for Stepwise” provides the ability to change the p-value according to how statistically significant the relationship between the X and Y-variables should be. The lower the p-value, the more significant the relation is.

**Sampling**

The “Sampling” slider adjusts how many records to use to create the model, and how many to hold out from development to evaluate a completed model’s accuracy. Records are chosen at random.

**Modeling Options**

**Include Intercept**

The intercept variable represents an inherent "bias" in the dataset, skewing the predictive results in a particular direction. The option will show or not show the Intercept in the final list of model variables.

**Use Forward/Backward Elimination**

By checking “Use forward / backward elimination”, Analytics will check to see if any of the variables should be dropped from the model during every model-building step. If a variable that has already entered the model is no longer statistically significant at the chosen p-value, that variable will be removed from the model.
Use Fast Model Method

Checking the "Use fast model method" box speeds up model development time by eliminating any variables once their significance level falls below the chosen threshold. All variables will be considered during each step if this option is not checked. Automatically-developed models may also take significantly longer to build on large datasets.

Use Parallel Algorithms

Using Parallel Algorithms utilizes multiple CPU cores when available in order to speed up model generation.

Use Previous Coefficients Optimization

By checking the “Use Previous Coefficients Optimization,” Analytics sets the initial regression coefficients to the coefficients computed at the previous iteration of the algorithm. This will speed up the model generation, though may produce a slightly different model.

Use Sparse Matrix Optimization

If the source data set is large and contains mostly 0’s or null values, or has categorical variables with many values, checking this option may provide a significant speedup in model generation.

Show Standardized Estimates

By checking the “Show standardized estimates” box, the output will include a column labeled “Std Est”. This column represents the estimates of the regression coefficients from the “Coef” column standardized to a normal distribution. This option is only available when working with an OLS Regression model.

Show Multicollinearity Diagnostics

Checking the “Show Multicollinearity diagnostics” box displays any highly correlated relationships between predictor variables. Although correlated variables can exist in a successful predictive model, the values assigned to their coefficients may not accurately estimate each collinear variable’s impact on the Y-variable if the other correlated variables were to be removed. Diagnostics we can include in our output are:

- **Variance Inflation Factor & Tolerance**: The Variance Inflation Factor (VIF) is equal to the reciprocal of tolerance. In general, the VIF represents the severity of Multicollinearity by measuring how much the variance of a coefficient is increased because of collinearity.
• **Covariance of the Estimates**: The matrix for the covariance of the estimates can be found in the “Diagnostics” section of the output. In general, a large positive value of covariance can mean that two variables are linearly related; a negative value can indicate that the two variables tend to move in opposite directions.

• **Correlation of the Estimates**: The correlation matrix can be found in the “Diagnostics” section of the output. The correlation between two estimates is equal to the slope of the regression line between the variables when they have both been normalized. Thus, a value of one indicates a direct correlation, and the closer to +1 or -1 the value is, the more correlated the two variables are. The sign of the correlation represents a positive or inverse relationship.

• **Correlation of the Variables**: The correlation matrix can be found in the “Diagnostics” section of the output. The correlation between two variables is equal to the slope of the regression line between the two variables. Thus, a value of one indicates a direct correlation, and the closer to +1 or -1 the value is, the more correlated the two variables are. The sign of the correlation represents a positive or inverse relationship.

• **x’x Matrix**: The x’x matrix (also called the variance-covariance matrix) can be found in the “Diagnostics” section of the output. This matrix displays the variances of each variable on the diagonal and the covariances between variables as the off-diagonal values.

**Use Weighted Regression**

By checking the “Use weighted regression” box and filling in the “Select Weight Variable” field with a continuous variable, we opt to assign a weight to each row equal to the value of the weight variable for that row. A weight value of zero would, therefore, cause a row to be ignored, and a weight value of two would cause that row to be weighted twice – as though that row appeared twice in the dataset.

**Quantiles**

The graphical representation of the model's self test may be changed using this control. Though typically grouping the test records into ten equal bins, this may be adjusted to increase or decrease the number of organizing bins.

**Variables Sorting**

The list of variables that appear in the model may be ordered either alphabetically in ascending or descending order, or based upon the reliability of the variable's influence on the Y-variable.
Decision Tree (CHAID) Modeling Options

Unlike the other modeling methods within Analytics, decision tree modeling must be actively opted into. This is done by opening the TOOLS > OPTIONS menu and under the modeling tab checking the "Enable Decision Tree (CHAID) model type" box. Once this is done, the Decision Tree regression type option will appear.

Selecting the Decision Tree (CHAID)" from the regression types section will change the view of options available as well as re-configuring the main Model screen to facilitate decision tree model development and presentation.

Missing Values Handling
The Missing Handling column identifies any missing values within each variable. If missing values are found, an option must be selected from the drop-down menu. “Default” (which is defined on the “Control Panel” tab), “Fill with Variable Mean”, or “Exclude Entire Record”. It is important to keep in mind that excluding multiple records can greatly reduce sample size for analysis; do so with caution.

Fill with Variable Means
Clicking the “Fill with Variable Means” button will fill in the missing values for each variable with the average value for that variable.

Exclude Entire Record
Clicking the “Exclude Entire Record” will not replace a missing value with any other value; rather, it will throw out that record for all variables so that it will not affect the analysis.
**Decision Tree Options**

**Alpha Merge**
This setting represents the minimum p-value required to merge two categories.

**Alpha Split**
This setting represents the maximum p-value required to split a category.

**Alpha Split-Merge**
This setting represents the maximum p-value required to perform a binary split a category after merging three or more original categories.

**Min Segment Size**
Minimum number of training samples required to split a node.

**Max Initial Ordinals**
This setting represents the maximum number of ordinals to split a continuous predictor variable

**Max Tree Depth**
This setting represents the maximum depth (number of levels) of the decision tree.

**Sampling**
The "Sampling" slider adjusts how many records to use to create the model, and how many to hold out from development to evaluate a completed model's accuracy. Records are chosen at random.

**Quantiles**
The graphical representation of the model's self test may be changed using this control. Though typically grouping the test records into ten equal bins, this may be adjusted to increase or decrease the number of organizing bins.

**Variables Sorting**
The list of variables that appear in the model may be ordered either alphabetically in ascending or descending order, or based upon the reliability of the variable's influence on the Y-variable.
Variables List

This is a listing of all the variables originally found in the dataset, as well as those that may have been created as part of the analysis, that may legitimately be considered for inclusion in the model. Variables that were manually excluded during the analysis and those that are automatically disqualified (constants, dates, indexes) will not appear.

This section acts as a pick list as the model is constructed.

Included Variables List

Variables that are manually selected from the main Variables List or those automatically selected by the Build Automatically process are listed here. These constitute the variables that will make up the model if it is then memorized.
Users may choose to build their own models by manually adding and/or removing variables or use the same method to change a “Build Model Automatically” model. Clicking “Build” will then regressively analyze the customizations and produce the model.

Clicking the “Suggest Variable” button allows Analytics to decide which variable from the “Variable List” window would best improve the model and moves it automatically to the “Include these” window.

By clicking the “Build Stepwise” button, Rapid Insight Analytics will check to see if any of the variables should be dropped from the model following each step of the regression analysis. If a variable that has already entered the model is no longer statistically significant at the chosen p-value, that variable will be removed from the model.
Build Automatically

The easiest and fastest way to build a model is by clicking the “Build Model Automatically” button. This option allows Rapid Insight Analytics to build what it feels is the best model that can be built on the current dataset. It takes all non-linearities in the data into account and makes all of the appropriate transformations on the candidate variables automatically.

Decision Trees (CHAID)

Decision Tree modeling is a method of predictive modeling that provides its results in a hierarchic fashion.

Enabling Decision Tree Modeling

Unlike the other modeling methods within Analytics, decision tree modeling must be actively opted into. This is done by opening the TOOLS > OPTIONS menu and under the modeling tab checking the "Enable Decision Tree (CHAID) model type" box.
Once this is done, a new “Decision Trees” Options tab appears. This tab allows the user to configure the default modeling parameters.

Defining the Conditions for the Decision Tree Model

To begin building a decision tree model, the user must open the Model Options sidebar and select Decision Tree (CHAID)” from the regression types section. This will change the view of options available as well as re-configuring the main Model screen to facilitate decision tree model development and presentation.

The choices available on this sidebar view are explained in detail in the Modeling Options help section.
Creating a Decision Tree Model

Return to the Model Building tab and begin selecting the variables that will participate in the modeling process. This is done by individually highlighting an entry in the "Variables" list on the left and moving it using the arrow controls to the "Included Variables" list on the right. Once all the desired variables are selected, click the Build button to generate the decision tree model.

The model's hierarchic structure begins collapsed under the Final Regression Model section. Clicking the arrow control beside the "All Samples" entry will begin to expand the model by level (node). Wherever a similar arrow control appears, another decision tree level/node can be revealed. The number of levels/nodes and the range of unique entries (branches) at each level/node are determined by the options selected under Model Options sidebar and the number and nature of the included variables chosen.

The "mu" value beside each entry represents the probability of the chosen Y-variable occurring based on the combination of variables represented by that part of the tree. The higher the value, the more likely the outcome given those factors are in place.
Final Regression Model Frame

Final Regression Model Tab

By selecting the Final Regression Model tab, once you’ve built an OLS or Logistic model, the user receives a detailed description of the current model. Descriptive sections include:

- Predicting
- Diagnostics
- Variable Contribution

Predicting Section

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coef</th>
<th>S.E.</th>
<th>Wald chi-sqr</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>12.12</td>
<td>2.513</td>
<td>23.27</td>
<td>0.000001</td>
</tr>
<tr>
<td>In_State</td>
<td>1.036</td>
<td>0.05775</td>
<td>321.91</td>
<td>0.000000</td>
</tr>
<tr>
<td>Square(SAT_Verbal)</td>
<td>-0.000014</td>
<td>0.0000003</td>
<td>17.01</td>
<td>0.000037</td>
</tr>
<tr>
<td>Athlete</td>
<td>1.347</td>
<td>0.2134</td>
<td>39.81</td>
<td>0.000000</td>
</tr>
<tr>
<td>SAT_Math</td>
<td>-0.00231</td>
<td>0.000447</td>
<td>26.89</td>
<td>0.000000</td>
</tr>
<tr>
<td>Binary(FATSAFiled,1)</td>
<td>0.3391</td>
<td>0.06006</td>
<td>31.23</td>
<td>0.000000</td>
</tr>
<tr>
<td>Binary(DEPT_CODE, STAT)</td>
<td>1.041</td>
<td>0.3055</td>
<td>12.00</td>
<td>0.000533</td>
</tr>
<tr>
<td>Binary(WEB_APPL_FLAG,1)</td>
<td>0.2292</td>
<td>0.06066</td>
<td>12.96</td>
<td>0.000318</td>
</tr>
<tr>
<td>SAT_Verbal</td>
<td>0.01318</td>
<td>0.00383</td>
<td>11.87</td>
<td>0.000570</td>
</tr>
<tr>
<td>LOG(Days Between App and Term Start)</td>
<td>-2.964</td>
<td>0.4703</td>
<td>39.72</td>
<td>0.000000</td>
</tr>
<tr>
<td>Days Between App and Term Start</td>
<td>0.00714</td>
<td>0.00118</td>
<td>36.49</td>
<td>0.000000</td>
</tr>
<tr>
<td>Binary(DEPT_CODE, POLY)</td>
<td>-0.5041</td>
<td>0.1986</td>
<td>6.442</td>
<td>0.01115</td>
</tr>
<tr>
<td>Binary(DEPT_CODE, CMN)</td>
<td>-0.4129</td>
<td>0.1709</td>
<td>5.333</td>
<td>0.01573</td>
</tr>
<tr>
<td>Square(HS_PERCENTILE)</td>
<td>-0.000033</td>
<td>0.000016</td>
<td>4.433</td>
<td>0.03524</td>
</tr>
</tbody>
</table>

The “Predicting” section appears at the top of the Final Regression Model section. It displays the variables that constitute the model, along with their coefficients, standard errors, Wald chi-square t-values, and p-values.
Diagnostics Section

Scrolling down the Final Regression Model section, the “Diagnostics” section provides objective measures evaluating the quality of the model. These measures include:

- **Intercept and Covariates values**
- **Odds Ratio Estimates** for each variable in the model
- 95% Confidence Limits (lower and upper) for each variable in the model
- **Percent Concordance**
- **Percent Discordance**
- Percent Tied
- Cox and Snell Pseudo R-squared score
- **Somers’ D score**
- **G-K Gamma score**
- **Kendall's Tau-a score**
- **C value**
Variable Contribution Section

The Variable Contribution section appears at the bottom of the Final Regression Model frame. It lists the variables that constitute the current model in order of their relative contribution to the overall probability score.

Model Steps Tab

Analytics will check to see if any of the variables should be dropped from the model following each step of the regression analysis. All of the regression analysis steps taken to obtain the current model are available by selecting and expanding the Model Steps section.

If a variable that has already entered the model is no longer statistically significant at the chosen p-value, that variable will be removed from the model.
Model Visualization Frame

The Model Visualization frame offers graphical representations of the results of the model having been tested by being applied to the holdout sample records. The red bars depict the actual behavior found in the sample. A set blue bars depicting the predicted outcome will appear when the “Show Actual and Predicted” option box is checked. There are four separate views - (Decile) Analysis, (Decile) Lift, Cumulative Lift and ROC Curve.

Decile Lift

The decile lift shows the percentage increase in accurate prediction by using the given model.
Decile Analysis

The decile analysis divides the total population into ten equal bins (deciles). The bars on this chart represent the average probability score for the records placed in each bin.

Cumulative Lift

The cumulative lift shows the cumulative percentage increase in accurate prediction for the given decile and lower.
ROC Curve Tab

The “ROC Threshold” represents the threshold at which you consider a record to be a 1 vs. a 0. In other words, when the threshold is at .5, all those records receiving a calculated probability exceeding .5 will be considered 1’s, while those below will be considered 0’s.

Setting a higher threshold will be a more conservative cutoff point, so that, for instance, only those with a probability higher than .75 will be considered 1’s. This is more conservative because it assumes less risk in assigning “positives” or 1’s.

Setting a lower threshold accomplishes the opposite.

Now, the ROC curve itself represents the C statistic for the model in question.
Moving the slider at the bottom of the pane adjusts the cutoff, and based on your chosen threshold, as well as the ROC chart for your model, it will report back your “True Positive” and “False Positive” rates. A “True Positive” is a record that has been assigned a “1” (based on the predicted probability and the selected threshold) and actually was a “1” historically. A “False Positive” is a record that has been assigned a “1” (again, based on the predicted probability and the selected threshold) and historically, was actually a “0.” You would want a high “True Positive” rate, and a low “False Positive” rate.

The basic way to discern whether the ROC Curve is indicating your model is good or bad depends on the shape of the curve. A “flatter” curve indicates that you cannot lower the ROC Threshold without increasing the “False Positive” rate significantly. This is not a desirable characteristic of a model.

True Positive: 0.48
False Positive: 0.20

A steeper curve (one that tracks closer to the top left corner) indicates that you can choose a lower ROC Threshold without increasing your "False Positive" rate significantly. This is a desirable characteristic of a model.
Report Screen

Add to Report

The ‘Add To Report’ button appends the chart or graph in view to Analytics’ built-in reporting feature. This is available within the Visualize Tab, the Clustering Tab and the Model Tab. To view the report itself, click on the "Report" Tab.

Report Screen

The Report Screen provides a handy place to incorporate any charts into one of several available reporting formats. When all of the desired charts have been assembled, clicking on the "Generate Report" ( ) control will publish the contents as a web page (.html), a slide show (.pptx), a spreadsheet (.xlsx) or a document (.docx or .pdf). The contents will remain in the Report Screen as long as the analysis is active or until removed using one of the delete slide controls ( ) at the top of the frame.

Individual graphics may individually viewed, annotated and saved (as a .png) by highlighting the desired graphic and clicking the "Open Selected Slide" ( ) control at the top of the frame.
Saving the Model

Memorize Model

To capture a model, enter a name in the “Model Name” field and click the “Memorize Model” button. This retains the model for recall later if the analysis is saved. It also makes it available for other actions in the program, such as Save Model(s), Create Model Variable and What If analysis.

What If

The “What If?” button allows users to explore how the Y-variable would be affected if one or more of the related variables were increased or decreased.

Note: ‘What If’ functionality makes assumptions regarding causalities in the data and for this reason results could turn out to be much different than projected. Rapid Insight® makes no guarantees regarding accuracy of these projections either implied or explicit and this tool is to be used at your own risk.

SELECT A MODEL / MODEL VARIABLES

Explore “What If?” scenarios using any memorized model by selecting it from the “Memorized Models” list. The “Model Equation” window displays the established formula for the selected model. To perform a “What If?” analysis, select any of the variables in the “Model Variables” window and use the slider control to change its value. The start value for any variable is its mean. The graph of the variable to the right will change to reflect the variable’s “new” mean as one or more a slider controls is repositioned.
Create Model Variable

Sometimes it is advantageous to view the probability scores a model produces as if it were a data element in the dataset under analysis. By selecting a memorized model and clicking the Create Model Variable icon, a new column is added to the list of variables - "Probability of Model: <name of model>". In effect, Analytics uses the selected model to score each row in the original dataset and stores the resulting probability value in this new column.

Save Models

In order to score a model later one, save it here by selecting just one model from the list of memorized models then clicking the black “Save Model(s)” icon. A file with a “.rism” file type is produced.
Report Screen

Add to Report

The 'Add To Report' button appends the chart or graph in view to Analytics' built-in reporting feature. This is available within the Visualize Tab, the Clustering Tab and the Model Tab. To view the report itself, click on the "Report" Tab.

Report Screen

The Report Tab provides a handy place to incorporate any charts into one of several available reporting formats. When all of the desired charts have been assembled, clicking on the "Generate Report" ( ) control will publish the contents as a web page (.html), a slide show (.pptx), a spreadsheet (.xlsx) or a document (.docx or .pdf). The contents will remain in the Report Tab as long as the analysis is active or until removed using one of the delete slide controls ( ) at the top of the frame.

Individual graphics may individually viewed, annotated and saved (as a .png) by highlighting the desired graphic and clicking the "Open Selected Slide" ( ) control at the top of the frame.
SCORING

Score a File

The Score a File button is found on the Workspace tab. It provides users a way within the product to apply saved and/or memorized models to new datasets and score the records. The scoring process results in a probability score being calculated for each row in the selected dataset.

OPEN FILE

When the "Score a File" button is clicked, the "Open File" dialog box opens. This is where the dataset to be scored is selected. Once a file is selected, a new tab ("Score") will open in Analytics.
The "Score" tab is where the process of scoring a dataset takes place. The columns of the selected dataset list on the left side of the screen. The user can choose to either employ a previously saved model ("Load Scoring Model") or make use of a model that has been memorized in an open analysis ("Select Memorized Model").

Click “Load Scoring Model” to select a previously saved Rapid Insight Scoring Model (.rism) file for use during the scoring process. If successful, the model will
SELECT MEMORIZED MODEL

Click “Select Memorized Model” to choose one of the predictive models that has been memorized in one of the open analyses for use during the scoring process.

OUTPUT OPTIONS

Once a model has been selected for use (either a saved or memorized) the model equation will appear in the window. If columns are present in the dataset being scored that the model does not require, these will be highlighted in green in the Columns list. If columns that the model requires are missing from the dataset, these will be listed in red.
Score the Table and Validate the Model

If a column corresponding to the model's Y-variable appears in the dataset, the output option “Score the table and validate the model” will be available. Choosing “Score the dataset and validate the model” will generate a chart below the drop-down menu displaying the decile analysis or cumulative lift (decided by the toggle buttons to the right of the drop-down menu) of the model for predicting the given Y-variable when the dataset is scored.

Just Score the Table

Choosing the “Just score the table” will just list the number of records, number of variables and whether the scoring process was successful (or not).
START SCORING

In all cases, when the "Start Scoring" button is clicked, an output file is produced. It is a duplicate of the original scoring file with an added column ("Predicted Y") where the probability score calculated for each row is saved.

Selecting one of the quantiling options under the drop down control “Don’t Output Decile Info” adds a second column to the output file listing the partition that each observation falls under.
### APPENDIX A - TASK FUNCTIONS SUMMARY

#### Function Descriptions (Multi-Variable Transform Task)

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abs</td>
<td>ABS( )</td>
<td>Returns the absolute, positive value of the given numeric expression</td>
</tr>
<tr>
<td>ACos</td>
<td>ACOS( )</td>
<td>Returns the angle, in radians, whose cosine is the given real expression</td>
</tr>
<tr>
<td>ASin</td>
<td>ASIN( )</td>
<td>Returns the angle, in radians, whose sine is the given real expression</td>
</tr>
<tr>
<td>ATan</td>
<td>ATAN( )</td>
<td>Returns the angle in radians whose tangent is the given real expression</td>
</tr>
<tr>
<td>ATn2</td>
<td>ATN2( #y, #x )</td>
<td>Returns the angle, in radians, whose tangent is the quotient (Y/X) of given real expressions</td>
</tr>
<tr>
<td>Average</td>
<td>AVERAGE( P1, expr, ... )</td>
<td>Returns the average value of a list of 1 or more numeric values.</td>
</tr>
<tr>
<td>Ceiling</td>
<td>CEILING( )</td>
<td>Returns the smallest integer, greater than, or equal to, the given numeric expression</td>
</tr>
<tr>
<td>Char</td>
<td>CHAR( )</td>
<td>A string function that converts an int ASCII code to a character</td>
</tr>
<tr>
<td>CharIndex</td>
<td>CHARINDEX( 'string', )</td>
<td>Returns the starting position of the specified expression in a character string. Returns 0 if not found.</td>
</tr>
<tr>
<td>Coalesce</td>
<td>COALESCE( P1, expr, ... )</td>
<td>Returns the first non-null value in a list of 2 or more values.</td>
</tr>
<tr>
<td>Contains</td>
<td>CONTAINS( 'string', )</td>
<td>Returns 1 if the specified expression is in the character string, otherwise returns 0</td>
</tr>
<tr>
<td>Cos</td>
<td>COS( )</td>
<td>Returns the trigonometric cosine of the given angle (in radians) in the given expression</td>
</tr>
<tr>
<td>Cot</td>
<td>COT( )</td>
<td>Returns the trigonometric cotangent of the specified angle (in radians) in the given real expression</td>
</tr>
<tr>
<td>Date</td>
<td>DATE( &quot;&quot; )</td>
<td>Converts a value to a date.</td>
</tr>
<tr>
<td>DateFirstOfMonth</td>
<td>DATEFIRSTOFMONTH( )</td>
<td>Determines the date of the first day of the month.</td>
</tr>
<tr>
<td>DateFirstOfMonth</td>
<td>DATEFIRSTOFMONTH( #year, #month )</td>
<td>Determines the date of the first day of the month.</td>
</tr>
<tr>
<td>DateLastOfMonth</td>
<td>DATELASTOFMONTH( )</td>
<td>Determines the date of the last day of the month.</td>
</tr>
<tr>
<td>Function</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>----------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td><code>DATELASTOFMONTH( #year, #month )</code></td>
<td>Determines the date of the last day of the month.</td>
<td></td>
</tr>
<tr>
<td><code>DATESTR( )</code></td>
<td>Convert a date to a string</td>
<td></td>
</tr>
<tr>
<td><code>DATESTR( , 'd' )</code></td>
<td>Formats a date as a string</td>
<td></td>
</tr>
<tr>
<td><code>DAY( )</code></td>
<td>Returns an integer representing the day datepart of the specified date (1-31)</td>
<td></td>
</tr>
<tr>
<td><code>DAYOFWEEK( )</code></td>
<td>Returns a string representing the day datepart of the specified date (Sunday, Monday, etc.)</td>
<td></td>
</tr>
<tr>
<td><code>DAYOFWEEKN( )</code></td>
<td>Returns an integer representing the day datepart of the specified date (1-7)</td>
<td></td>
</tr>
<tr>
<td><code>DAYOFYEAR( )</code></td>
<td>Returns an integer representing the day datepart of the specified date (1-366)</td>
<td></td>
</tr>
<tr>
<td><code>DAYSBETWEEN( , date )</code></td>
<td>Returns an integer representing the # of days between 2 dates</td>
<td></td>
</tr>
<tr>
<td><code>DEGREES( )</code></td>
<td>Given an angle in radians, returns the corresponding angle in degrees</td>
<td></td>
</tr>
<tr>
<td><code>DISTANCEBETWEEN( , zipcode )</code></td>
<td>Returns a real representing the # of miles between 2 zipcodes</td>
<td></td>
</tr>
<tr>
<td><code>DISTANCEBETWEENLATLON( , lon1, lat2, lon2 )</code></td>
<td>Returns a real representing the # of miles between 2 Latitude/Longitude coordinate pairs.</td>
<td></td>
</tr>
<tr>
<td><code>DISTANCEBETWEENMGRS( , mgrsCoord )</code></td>
<td>Returns a integer representing the # of meters between 2 MGRS coordinates (in same GZD).</td>
<td></td>
</tr>
<tr>
<td><code>DISTANCEFROM( , 'zipcode' )</code></td>
<td>Returns a real representing the # of miles between a zipcode and a constant zip code</td>
<td></td>
</tr>
<tr>
<td><code>DISTANCEFROMLATLON( , varLon1, constLat2, constLon2 )</code></td>
<td>Returns a real representing the # of miles between a variable and a constant Latitude/Longitude coordinate pair.</td>
<td></td>
</tr>
<tr>
<td><code>DISTANCEFROMMGRS( , 'mgrsCoord' )</code></td>
<td>Returns a real representing the # of meters between a MGRS coordinate and a constant MGRS coordinate (in same GZD).</td>
<td></td>
</tr>
<tr>
<td><code>EXP( )</code></td>
<td>Returns the exponential value of the given real expression</td>
<td></td>
</tr>
<tr>
<td><code>FLOOR( )</code></td>
<td>Returns the largest integer less than or equal to the given numeric expression</td>
<td></td>
</tr>
<tr>
<td><code>GREATEST( , expr, ... )</code></td>
<td>Returns the largest value in a list of 2 or more values</td>
<td></td>
</tr>
<tr>
<td><code>GREATESTINDEX( , expr, ... )</code></td>
<td>Returns the index of the largest value in a list of 2 or more values.</td>
<td></td>
</tr>
<tr>
<td><code>HOUR( )</code></td>
<td>Returns an integer that represents the hour part of a specified datetime</td>
<td></td>
</tr>
<tr>
<td><code>HOURSBETWEEN( , date )</code></td>
<td>Returns an integer representing the # of hours between 2 dates</td>
<td></td>
</tr>
<tr>
<td><code>IF( , trueValue, falseValue )</code></td>
<td>Tests an expression result, returns one value if TRUE and another value if FALSE</td>
<td></td>
</tr>
<tr>
<td><code>ISNULL( , replacement_value )</code></td>
<td>Tests if an expression result is null, returns the result if it is not null, otherwise a replacement value</td>
<td></td>
</tr>
<tr>
<td><code>IS NULL ( IS NULL)</code></td>
<td>Tests if an expression result is null, evaluates to boolean</td>
<td></td>
</tr>
<tr>
<td><code>LAG( , expr )</code></td>
<td>Returns the prior value of the expression.</td>
<td></td>
</tr>
<tr>
<td><code>LAG( , expr )</code></td>
<td>Returns a value of the expression from P2 rows ago.</td>
<td></td>
</tr>
<tr>
<td><code>LEAST( , expr, ... )</code></td>
<td>Returns the smallest value in a list of 2 or more values</td>
<td></td>
</tr>
<tr>
<td><code>LEFT( , #len )</code></td>
<td>Returns the left part of a string the specified number of characters in length</td>
<td></td>
</tr>
<tr>
<td><code>LEN( )</code></td>
<td>Returns the number of characters of the given string, excluding trailing blanks</td>
<td></td>
</tr>
<tr>
<td><code>LOG( )</code></td>
<td>Returns the natural logarithm of the given real expression</td>
<td></td>
</tr>
<tr>
<td><code>LOG10( )</code></td>
<td>Returns the base-10 logarithm of the given real expression</td>
<td></td>
</tr>
<tr>
<td>Function</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>-----------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Lower</td>
<td><code>LOWER( )</code> Returns a character expression after converting uppercase character data to lowercase</td>
<td></td>
</tr>
<tr>
<td>LTrim</td>
<td><code>LTRIM( )</code> Returns a character expression after removing leading blanks</td>
<td></td>
</tr>
<tr>
<td>MatchesRegEx</td>
<td><code>MATCHESREGESEX( , 'regex')</code> Tests if an expression result matches a regular expression. (See TRANSFORM for additional info.)</td>
<td></td>
</tr>
<tr>
<td>MetersToMiles</td>
<td><code>METERSTOMILES( )</code> Returns the real number of miles equivalent to the specified number of meters.</td>
<td></td>
</tr>
<tr>
<td>Minute</td>
<td><code>MINUTE( )</code> Returns an integer that represents the hour part of a specified datetime</td>
<td></td>
</tr>
<tr>
<td>MinutesBetween</td>
<td><code>MINUTESBETWEEN( , date )</code> Returns an integer representing the # of minutes between 2 dates</td>
<td></td>
</tr>
<tr>
<td>Month</td>
<td><code>MONTH( )</code> Returns an integer that represents the month part of a specified date (1-12)</td>
<td></td>
</tr>
<tr>
<td>MonthOfYear</td>
<td><code>MONTHOFYEAR( )</code> Returns a string that represents the month part of a specified date (January, February, etc.)</td>
<td></td>
</tr>
<tr>
<td>MonthsBetween</td>
<td><code>MONTHSBETWEEN( , date )</code> Returns an integer representing the # of month boundaries between 2 dates.</td>
<td></td>
</tr>
<tr>
<td>MonthsBetween</td>
<td><code>MONTHSBETWEEN( , date, 1 )</code> Returns an integer representing the absolute # of months between 2 dates.</td>
<td></td>
</tr>
<tr>
<td>Now</td>
<td><code>NOW()</code> Returns the current date/time</td>
<td></td>
</tr>
<tr>
<td>NullIf</td>
<td><code>NULLIF( , expr )</code> Returns a null if two expressions are equivalent</td>
<td></td>
</tr>
<tr>
<td>Num</td>
<td><code>NUM( )</code> Returns numeric data converted from character data</td>
<td></td>
</tr>
<tr>
<td>PadLeft</td>
<td><code>PADLEFT( , #len, '' )</code> Returns a string padded on the left with specified character to the specified length</td>
<td></td>
</tr>
<tr>
<td>PadRight</td>
<td><code>PADRIGHT( , #len, '' )</code> Returns a string padded on the right with specified character to the specified length</td>
<td></td>
</tr>
<tr>
<td>PI</td>
<td><code>PI()</code> Returns the constant value of PI (3.14...)</td>
<td></td>
</tr>
<tr>
<td>Power</td>
<td><code>POWER( , #pow )</code> Returns the value of the given expression to the specified power</td>
<td></td>
</tr>
<tr>
<td>Quarter</td>
<td><code>QUARTER( )</code> Returns an integer representing the 'Quarter' specified date falls within.</td>
<td></td>
</tr>
<tr>
<td>Radians</td>
<td><code>RADIANS( )</code> Returns radians when a numeric expression, in degrees, is entered</td>
<td></td>
</tr>
<tr>
<td>Random</td>
<td><code>RANDOM()</code> Returns a random number between 0.0 and 1.0</td>
<td></td>
</tr>
<tr>
<td>Replace</td>
<td><code>REPLACE( , 'find', 'new' )</code> Replaces all occurrences of the first string expression with the second string expression</td>
<td></td>
</tr>
<tr>
<td>RemoveLeft</td>
<td><code>REMOVEDLEFT( , #cnt )</code> Removes specified number of characters from left end of string expression</td>
<td></td>
</tr>
<tr>
<td>RemoveRight</td>
<td><code>REMOVEDRIGHT( , #cnt )</code> Removes specified number of characters from right end of string expression</td>
<td></td>
</tr>
<tr>
<td>Reverse</td>
<td><code>REVERSE( )</code> Returns the reverse of a character expression</td>
<td></td>
</tr>
<tr>
<td>Right</td>
<td><code>RIGHT( , #len )</code> Returns the right part of a string the specified number of characters in length</td>
<td></td>
</tr>
<tr>
<td>Round</td>
<td><code>ROUND( )</code> Round value to the nearest integer value</td>
<td></td>
</tr>
<tr>
<td>RTrim</td>
<td><code>RTRIM( )</code> Returns a character string after truncating all trailing blanks</td>
<td></td>
</tr>
<tr>
<td>Second</td>
<td><code>SECOND( )</code> Returns an integer that represents the second part of a specified datetime</td>
<td></td>
</tr>
<tr>
<td>SecondsBetween</td>
<td><code>SECONDSBETWEEN( , date )</code> Returns an integer representing the # of seconds between 2 dates</td>
<td></td>
</tr>
<tr>
<td>Sign</td>
<td><code>SIGN( )</code> Returns the positive (+1), zero (0), or negative (-1) sign of the given expression</td>
<td></td>
</tr>
<tr>
<td>Sin</td>
<td><code>SIN( )</code> Returns the trigonometric sine of the given angle (in radians) in an approximate numeric (real) expression</td>
<td></td>
</tr>
<tr>
<td>Space</td>
<td><code>SPACE( )</code> Returns a string of repeated spaces</td>
<td></td>
</tr>
<tr>
<td>Square</td>
<td><code>SQUARE( )</code> Returns the square of a given expression</td>
<td></td>
</tr>
<tr>
<td>Sqrtr</td>
<td><code>SQRT( )</code> Returns the square root of the given expression</td>
<td></td>
</tr>
<tr>
<td>Function</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>-----------------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Str</td>
<td>Returns the character data converted from numeric data</td>
<td></td>
</tr>
<tr>
<td>Stuff</td>
<td>Deletes a specified length of characters and inserts another set of characters at a specified starting point</td>
<td></td>
</tr>
<tr>
<td>SubField</td>
<td>Returns the Nth 'sub-field' from a string where the fields are delimited by commas</td>
<td></td>
</tr>
<tr>
<td>SubField</td>
<td>Returns the Nth 'sub-field' from a string where the fields are delimited by the specified delimiter (default is comma)</td>
<td></td>
</tr>
<tr>
<td>SubFieldCount</td>
<td>Returns the number of 'sub-fields' in a string where the fields are delimited by a comma</td>
<td></td>
</tr>
<tr>
<td>SubFieldCount</td>
<td>Returns the number of 'sub-fields' in a string where the fields are delimited by the specified delimiter (default is comma)</td>
<td></td>
</tr>
<tr>
<td>SubString</td>
<td>Returns part of a string beginning at the specified start position to the end of the string</td>
<td></td>
</tr>
<tr>
<td>SubString</td>
<td>Returns part of a string beginning at the specified start position and continuing for the specified length</td>
<td></td>
</tr>
<tr>
<td>Tan</td>
<td>Returns the tangent of the input expression</td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>Converts a value to a time.</td>
<td></td>
</tr>
<tr>
<td>TitleCase</td>
<td>Returns a character expression after converting the character data to Title case.</td>
<td></td>
</tr>
<tr>
<td>Today</td>
<td>Returns the current date</td>
<td></td>
</tr>
<tr>
<td>Upper</td>
<td>Returns a character expression with lowercase character data converted to uppercase</td>
<td></td>
</tr>
<tr>
<td>Week</td>
<td>Returns an integer expression representing the week that the specified date falls in (1-53)</td>
<td></td>
</tr>
<tr>
<td>Weekday</td>
<td>Returns 'Weekday' or 'Weekend'</td>
<td></td>
</tr>
<tr>
<td>WorkdaysBetween</td>
<td>Returns an integer representing the # of workdays between 2 dates</td>
<td></td>
</tr>
<tr>
<td>Year</td>
<td>Returns an integer that represents the year part of a specified date</td>
<td></td>
</tr>
<tr>
<td>YearsBetween</td>
<td>Returns an integer representing the # of 12 month boundaries between 2 dates.</td>
<td></td>
</tr>
<tr>
<td>YearsBetween</td>
<td>Returns an integer representing the absolute # of years between 2 dates.</td>
<td></td>
</tr>
<tr>
<td>Zipcode</td>
<td>Returns a 5 character zipcode converted from string or numeric data. Illegal values are output as text</td>
<td></td>
</tr>
<tr>
<td>Zipcode</td>
<td>Returns a US zip or zip+4 converted from string or numeric data.</td>
<td></td>
</tr>
<tr>
<td>ZipcodeLatitude</td>
<td>Returns the latitude of a zipcode. Unrecognized zipcodes return missing value</td>
<td></td>
</tr>
<tr>
<td>ZipcodeLongitude</td>
<td>Returns a US zip or zip+4 converted from string or numeric data.</td>
<td></td>
</tr>
</tbody>
</table>
Function Operators

Because functions have such a high priority they almost always require their parameters to be placed in parentheses to achieve the expected/correct result.

"ABS [A]" is the same as "ABS ([A])", but "ABS [A]*[B]" is not "ABS ([A]*[B])", but is "ABS([A])*[B]"

<table>
<thead>
<tr>
<th>PRECEDENCE</th>
<th>SYMBOL</th>
<th>TYPE OF OPERATION</th>
<th>ASSOCIATIVELY</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>()</td>
<td>Expression</td>
<td>Left to right</td>
</tr>
<tr>
<td>2</td>
<td>-</td>
<td>Scalar Functions</td>
<td>Left to right</td>
</tr>
<tr>
<td>3</td>
<td>* / % **</td>
<td>Multiplicative</td>
<td>Left to right</td>
</tr>
<tr>
<td>4</td>
<td>+ -</td>
<td>Additive</td>
<td>Left to right</td>
</tr>
<tr>
<td>5</td>
<td>&lt; &gt;</td>
<td>Bitwise shift</td>
<td>Left to right</td>
</tr>
<tr>
<td>6</td>
<td>&lt;= &gt;=</td>
<td>Relational</td>
<td>Left to right</td>
</tr>
<tr>
<td>7</td>
<td>= == &lt;&gt; != IS</td>
<td>Equality</td>
<td>Left to right</td>
</tr>
<tr>
<td>8</td>
<td>&amp;</td>
<td>Bitwise-AND</td>
<td>Left to right</td>
</tr>
<tr>
<td>9</td>
<td>^</td>
<td>Bitwise-exclusive-OR</td>
<td>Left to right</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td>Bitwise-OR (inclusive)</td>
</tr>
<tr>
<td>11</td>
<td>BETWEEN</td>
<td>Logical-BETWEEN</td>
<td>Left to right</td>
</tr>
<tr>
<td>12</td>
<td>IN</td>
<td>Logical-IN</td>
<td>Left to right</td>
</tr>
<tr>
<td>13</td>
<td>LIKE</td>
<td>Logical-LIKE</td>
<td>Left to right</td>
</tr>
<tr>
<td>14</td>
<td>&amp;&amp; AND</td>
<td>Logical-AND</td>
<td>Left to right</td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td>OR</td>
</tr>
<tr>
<td>16</td>
<td>??</td>
<td>Null-Coalescing</td>
<td>Left to right</td>
</tr>
<tr>
<td>17</td>
<td>?::</td>
<td>Inline-If/Then</td>
<td>Left to right</td>
</tr>
<tr>
<td>18</td>
<td>,</td>
<td>List</td>
<td>Left to right</td>
</tr>
<tr>
<td>19</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Date and Time Format Specifiers

Many functions require or allow certain formatting options to be specified.

<table>
<thead>
<tr>
<th>SPECIFIER</th>
<th>DESCRIPTION</th>
<th>EXAMPLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;d&quot;</td>
<td>The day of the month, from 1 through 31.</td>
<td>6/1/2009 1:45:30 PM -&gt; 1</td>
</tr>
<tr>
<td>&quot;dd&quot;</td>
<td>The day of the month, from 01 through 31.</td>
<td>6/1/2009 1:45:30 PM -&gt; 01</td>
</tr>
<tr>
<td>&quot;ddd&quot;</td>
<td>The abbreviated name of the day of the week.</td>
<td>6/15/2009 1:45:30 PM -&gt; Monday (en-US)</td>
</tr>
<tr>
<td>&quot;dddd&quot;</td>
<td>The full name of the day of the week.</td>
<td>6/15/2009 1:45:30 PM -&gt; Monday (en-US)</td>
</tr>
<tr>
<td>&quot;f&quot;</td>
<td>The minutes offset from UTC.</td>
<td>6/15/2009 13:45:30.6175425</td>
</tr>
<tr>
<td>&quot;ff&quot;</td>
<td>The seconds offset from UTC.</td>
<td>6/15/2009 13:45:30.6175425</td>
</tr>
<tr>
<td>&quot;fff&quot;</td>
<td>The milliseconds in a date and time value.</td>
<td>6/15/2009 13:45:30.6175425</td>
</tr>
<tr>
<td>&quot;ffff&quot;</td>
<td>The hundredths of a second in a date and time value.</td>
<td>6/15/2009 13:45:30.6175425</td>
</tr>
<tr>
<td>&quot;ffffff&quot;</td>
<td>The hundredths of a second in a date and time value.</td>
<td>6/15/2009 13:45:30.6175425</td>
</tr>
<tr>
<td>&quot;fffffff&quot;</td>
<td>The milliseconds in a date and time value.</td>
<td>6/15/2009 13:45:30.6175425</td>
</tr>
<tr>
<td>&quot;yyyy&quot;</td>
<td>The year, from 00 to 99.</td>
<td>1/1/0001 12:00:00 AM -&gt; 001</td>
</tr>
<tr>
<td>&quot;yyyyy&quot;</td>
<td>The year as a four-digit number.</td>
<td>1/1/0001 12:00:00 AM -&gt; 0001</td>
</tr>
<tr>
<td>&quot;yyyyyy&quot;</td>
<td>The year as a five-digit number.</td>
<td>1/1/0001 12:00:00 AM -&gt; 00001</td>
</tr>
<tr>
<td>&quot;zz&quot;</td>
<td>Hours offset from UTC, with no leading zeros.</td>
<td>6/15/2009 13:45:30 PM -&gt; -7</td>
</tr>
<tr>
<td>&quot;zzz&quot;</td>
<td>Hours and minutes offset from UTC.</td>
<td>6/15/2009 13:45:30 PM -&gt; -07:00</td>
</tr>
<tr>
<td>&quot;s&quot;</td>
<td>The time separator.</td>
<td>6/15/2009 13:45:30 PM -&gt; : (en-US)</td>
</tr>
<tr>
<td>&quot;y&quot;</td>
<td>The date separator.</td>
<td>6/15/2009 13:45:30 PM -&gt; -6</td>
</tr>
</tbody>
</table>

GLOSSARY
## Special Parameters

This section documents unique parameters that do not fit neatly into other categories yet expand the user’s abilities and contribute to the utility of the program.

<table>
<thead>
<tr>
<th>SYMBOL</th>
<th>TYPE</th>
<th>BEHAVIOR</th>
<th>DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>*</td>
<td>Parameter</td>
<td>Wildcard</td>
<td>When an asterisk is used as a formula parameter, it acts as a wildcard. COMBINE INPUTS Table Pattern Example: <em><em>Enrollment Data</em>.</em>** Selects all files, regardless of type, that start with the characters &quot;Enrollment Data&quot;.</td>
</tr>
<tr>
<td></td>
<td>Functional Parameter</td>
<td>Wildcard</td>
<td>When the function is used as a formula parameter, it acts as a wildcard with features beyond those of an asterisk. COMBINE INPUTS Table Pattern Example: <strong>Regex:CX..\xls.<em>.</em>$</strong> Selects all worksheets from all Excel (.xls) files that start with the characters &quot;CX&quot; followed by exactly two characters.</td>
</tr>
<tr>
<td>[ ]</td>
<td>Parameter</td>
<td>Lag</td>
<td>An underscore in square brackets used as a formula parameter applies the value of the previous record. An example that creates an incrementing series in the new column starting at 100: <strong>IF ([ ] IS NULL, 100, [ ] + 1)</strong> The LAG () Transform function performs a similar function, but allows users to designate the column being lagged in a multivariable formula. It is recommended that if the function is complex, that the value of the lag be defined as a separate step: <strong>[prior]:= LAG([B]); IF([C]=7, [prior], [B])</strong></td>
</tr>
</tbody>
</table>

## Contact Us

For any issues/questions about Rapid Insight® Veera, please send an email to RI Support at support@rapidinsightinc.com
Glossary of Terms

Definition Akaike Information Criterion (AIC)

AIC is a measure of the relative quality of a statistical model for a given set of data. AIC deals with the trade-off between the goodness of fit of the model and the complexity of the model by offering a relative estimate of the information lost when a given model is used to represent the process that generates the data. The preferred model is the one with the minimum AIC value, however AIC can tell you nothing about the quality of the model in an absolute sense as it will not give any warning if all of the candidate models fit poorly.

Definition Automine

For categorical variables:

RI runs through each category of the variable and creates a binary 0/1 variable. For each category/binary transform, it finds it’s correlation with the Y variable and calculates the F-statistic for that category/binary transform. The F-statistic is a test used to evaluate whether or not the category/binary transform would be a contributor to the model. If a categorical variable contains at least one category that is found to be statistically significant, that variable is considered to be related to the Y. Binary variables are created for each category that is found to be a potential contributor to the model based on the F-statistic for that category. The correlations and F Values can be found on the automine screen, as well as in the visualization screen.

For continuous variables:

RI runs through several transforms for each continuous variable. It tests the natural log, the square root, the square, the cubed root, and the cube of the variable. It calculates the correlation with the Y as well as the F-statistic for each transform of the variable. If, based on the correlation and the F-statistic, the variable (or any transform of the variable) is found to be a potential contributor to the model, it considers that variable to be related to the Y.

Definition Binary Variable

A variable that uses only two possible categories, “0” or “1”.
Definition Categorical Variable

A variable where each observation can be classified as belonging to one of a finite number of possible categories or types (i.e. – Gender {male, female}). Categorical data may not be used as a Y-Variable.

Definition Concordance

Concordance is a means of evaluating a model’s predictive quality. It is only used when the Y-variable is binary. Each record in a dataset (where the Y-variable is known) is assigned a probability score by the model. Then, each record where the Y-variable is a “1” is paired with every record where the Y-variable is “0”. Every time the probability score of the record with a “1” is greater than the probability score of the record with a “0”, it is considered concordant. The ratio of concordant pairs to the total number of pairs is represented as a percentage.

Definition Continuous Variable

A continuous variable is one which takes on a non-countable infinite number of values.

Definition Correlation Analysis

Each variable that is part of the predictive modeling formula is assigned a coefficient number. This number is multiplied by the variable’s value to arrive at the variable’s contribution to a record’s probability score. A negative coefficient means the variable is inversely related to the Y-variable.

Definition Cumulative Lift

Cumulative Lift describes the advantage over random sampling from the whole dataset would be if you confined your attention to just those folks/records that are found in that decile and those deciles to the left of it. So if the cumulative lift of decile “4” is 35%, then, confining your interest to just those records in deciles 1,2,3 & 4 will provide you, on average, 35% better results than a random sampling of the whole dataset.

Definition Decile Analysis

One of the primary means by which Analytics evaluates the strength of a model. The program applies the model it has just developed to the hold-out records reserved for testing, calculating a probability score for each record. These records are then sorted using these scores before splitting them into 10 equal “deciles”. Decile 1 contains the 10% of records with the highest probability scores and decile 10 contains the lowest 10% of scores. Once this is done, then the average of the real outcomes is calculated for each decile and graphically displayed. The most desirable representation is a “staircase effect”, with the steps getting progressively lower from left to right.
Definition Delimiter

A data field delimiter is a character used to separate multiple columns in one row of a dataset.

Examples:

- Comma: CUST_ID,STREET,CITY,STATE,POSTCD, ZIP+4
- Tab: CUST_ID STREET CITY STATE POST CD ZIP+4
- Pipe: CUST_ID|STREET|CITY|STATE|POSTCD|ZIP+4

Definition Delta-P

“Delta-P” calculates the odds change of a one unit increase in that variable. The formula to convert an odds ratio estimate (show in Rapid Insight Analytics), to a Delta-P is simply:

\[
\text{Odds ratio estimate} / (1 + \text{odds ratio estimate}).
\]

Definition Discordance

Discordance is a means of evaluating a model’s predictive quality. It is only used when the Y-variable is binary. Each record in a dataset (where the Y-variable is known) is assigned a probability score by the model. Then, each record where the Y-variable is a “0” is paired with every record where the Y-variable is “1”. Every time the probability score of the record with a “0” is greater than the probability score of the record with a “1”, it is considered discordant. The ratio of discordant pairs to the total number of pairs is represented as a percentage.

Definition G-K Gamma

(“Goodman and Kruskal’s Gamma”) G-K Gamma reflects the difference of discordant pairs from the number of concordant pairs divided by the sum of non-tied pairs. The difference between this and Somer’s D is that tied pairs are not included in the denominator.

Definition Hold Out Sample

Hold-out Sample is used for validating the prediction accuracy on another sample. A hold-out sample (different records within the same dataset) is used because going out and getting completely fresh data is rarely done.
Definition Intercept

A predictive model's intercept is the coefficient that ensures that every line/plane/hyperplane does not have to be fit from a 0,0 coordinate. Basically it is there to ensure an optimal fit in the data.

Definition Kendall Tau A

Kendall Tau-A is a statistical measure of the strength of concordance and discordance (and so the predictive ability) of a model. It is presented as a number ranging between “1” (perfectly predictive) and “-1” (counter-predictive). A score of “0” signifies the model’s predictions are statistically random.

Definition Log Likelihood Function

Log Likelihood Function (or "-2 log l") is the natural logarithm of the Likelihood Function. The maximum likelihood estimate is the same for both the likelihood function and the natural logarithm of the likelihood function. It is generally used to find the maximum likelihood estimate.

Definition Logistic Regression

Logistic Regression is regression typically used to model a binary Y variable.

Definition Logs Ratio Estimate

The Logs Ratio Estimate explains what a one unit increase in that variable has on the probability of the outcome. A related measure, “Delta-P”, calculates the odds change of a one unit increase in that variable. The formula to convert an odds ratio estimate (show in Rapid Insight Analytics), to a Delta-P is simply:

\[
\text{Odds ratio estimate} / (1 + \text{odds ratio estimate}).
\]

Definition Mean

The mean of a variable is the sum of all the observations, divided by the number of observations.

Definition Model Formula Coefficient

Each variable that is part of the predictive modeling formula is assigned a coefficient number. This number is multiplied by the variable’s value to arrive at the variable’s contribution to a record’s probability score. A negative coefficient means the variable is inversely related to the Y-variable.
Definition Multivariate Analysis

The inter-relationships between two or three variables visually represented.

Definition Odds Ratio Estimate

Odds Ratio Estimate (ORE):

This is a means of conveying the impact a unit increase or decrease of the specific variable on the odds of a "success" observed in your y variable. This takes the baseline estimated probability into account while forming the "odds ratio." For instance, if the predicted probability of observing a "success" for a given record is 20%, the odds ratio is said to be the probability of "success" divided by the probability of a "non-success" in this case, \((0.20/0.80) = 0.25\) (the odds ratio).

Continuing the example, if the odds ratio estimate for the variable “GPA” is 4, a one unit increase in GPA (e.g. from 2.0 to 3.0) would quadruple the odds of a “success.” That is, the baseline odds ratio used in this example, of .25, would be multiplied by 4. The new odds ratio, then, is 1.

To determine the per-unit increase to probability, let \(x\) represent the new odds ratio after the unit increase or decrease. The result is the affected odds ratio, \(x\). To determine the new probability, calculate \(x/(1+x)\). This produces the new probability of a “success”.

C value:

This is the area under the ROC curve. It can be interpreted as a measure of model validity. It ranges from 0.5 to 1.0, and being closer to 1.0 is a desirable characteristic of a model.

Definition OLS Regression

Ordinary least squares (OLS) regression is used to model continuous Y variables. Regression models are powerful tools for predicting a score based on some other score. They involve a linear transformation of the predictor variable into the predicted variable. The parameters of the linear transformation are selected such that the least squares criterion is met, resulting in an "optimal" model. The model can then be used in the future to predict either exact scores, called point estimates, or intervals of scores, called interval estimates.

Definition Outlier

Outliers are values that are atypical, infrequent observations.
**Definition P-Value**

A p-value is a measure of statistical significance. The smaller the value, the more confidence can be placed in the relationship being reported.

To state it more formally, the p-value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true. One often "rejects the null hypothesis" when the p-value is less than the predetermined significance level, indicating that the observed result would be highly unlikely under the null hypothesis.

**Definition RIAX Filetype**

An RIAX file is the improved Rapid Insight Analysis file, new for version 3.0. It not only stores all the particulars of the chosen analysis, but it also saves all of the data upon which the analysis is based. This file may be loaded into Analytics to revisit and continue working with an earlier analysis.

**Definition RISM Filetype**

An RISM file is Rapid Insight Scoring Model file. It's a file that saves the selected predictive model. It can be used in conjunction with either Analytics "Scoring" feature or Veera's "Import Scoring Model" utility to ultimately score new datasets.

**Definition ROC Curve**

This is a chart which plots the ordered pairs (false positive rate, true positive rate) at each probability cutoff. This is only available when performing Logistic Regression (as opposed to OLS). The ordered pairs are determined by calculating the true and false positive rates at each available cutoff point. These cutoff points are the levels at which you can partition your predicted results into 1’s and 0’s (i.e. if cutoff point is .8, all records assigned a .8 or above are 1’s, and those below .8 are 0’s).

The area under the ROC curve is known a "C". It can be interpreted as a measure of model validity. It ranges from 0.5 to 1.0, and being closer to 1.0 is a desirable characteristic of a model.

**Definition Schwarz Criterion**

Schwarz Criterion (SC - also known as the Bayesian Information Criterion or BIC) is a criterion for model selection among a finite set of models. When fitting models it is possible to increase the likelihood by adding parameters, but doing so may result in overfitting. Both SC and AIC resolve this problem by introducing a penalty term for the number of parameters in the model.
Definition Somer's D

Somer's D reflects the difference of discordant pairs from the number of concordant pairs divided by the total number of pairs. Values range from -1 (all pairs disagree) to 1 (all pairs agree.) Larger values indicate that the model has better predictive ability. Primarily, this metric is used as a comparison tool between two different predictor variables or models.

Somer's D can be used as a model building criteria itself, applied to individual predictors, though it's most valuable in narrower analyses. In the end, however, Somer's D can also be applied to a scoring model as a whole. Then theoretically, if there is a question regarding which of two or more models you might use in the end (on the sample of records), the D statistic could be a deciding factor. There are some circumstances where a score of -1 could be desirable, but only when the D statistic is being used as the predictor selection criteria (a score of -1, after all, would represent a perfect negative predictor).

Definition Standard Deviation

Standard Deviation is the positive square root of the variance.

Definition Standard Error

Standard Error represents the standard deviation around the variable’s coefficient (see Coefficient) within which 68% of the records fell when the model was tested. The smaller the S.E. is in relation to the coefficient, the more accurate the model.

Definition Univariate Analysis

One variable is visually represented at a time.

Definition Wald Chi Square

Chi-Square sorts the data by the X-variable and creates 10 equal deciles based on the sorted data. It then runs a Chi-Square test on these deciles to determine whether the ‘ones’ are randomly distributed across the deciles. This is a more robust test than using a linear correlation, as it captures non-linear relationships, as well as relationships that are not well fit by a curve or a line.