# Tips, Tricks, and What’s New in IBM SPSS Modeler

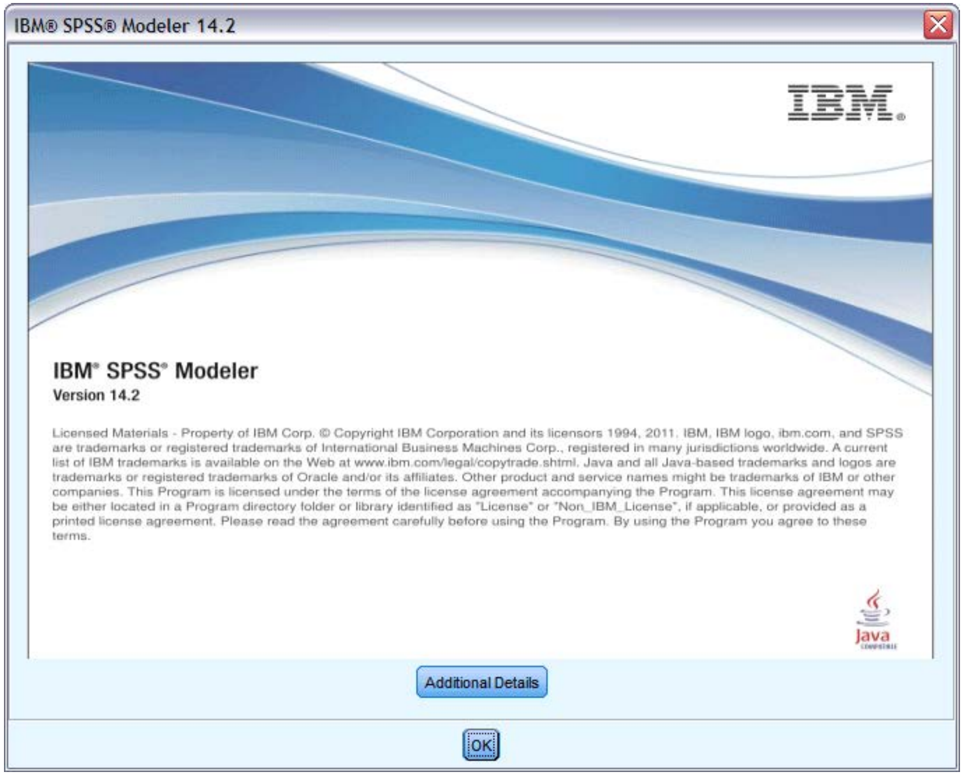
Business Analytics software

# Agenda, Goals for the Session, and Opening Remarks

- Welcome
- The Agenda for Today’s Session includes General Buckets Addressing:
  - Naming
  - GUI – the User Interface
  - Sources – Accessing Data
  - Data Preparation
  - Modeling
  - Integration
  - Output
- The Goal for Today’s Session
  - To take away at least one tip, trick, or fact about what’s new that will improve the efficiency and/or effectiveness of your data mining / predictive analytics efforts

# Intro: Clarifying the Name of IBM SPSS Modeler

- Version 12: “Clementine “
  - Long-time users refer to Modeler as “Clementine” or “Clem”
- Version 13: “PASW Modeler”
  - “PASW” was an umbrella term used to signify that Modeler was one facet of a Predictive Analytics Software portfolio solution (along side Data Collection, Statistics, Collaboration & Deployment Services…)
- Version 14: “IBM SPSS Modeler”
  - With the acquisition, IBM well-understood the equity of the SPSS brand name

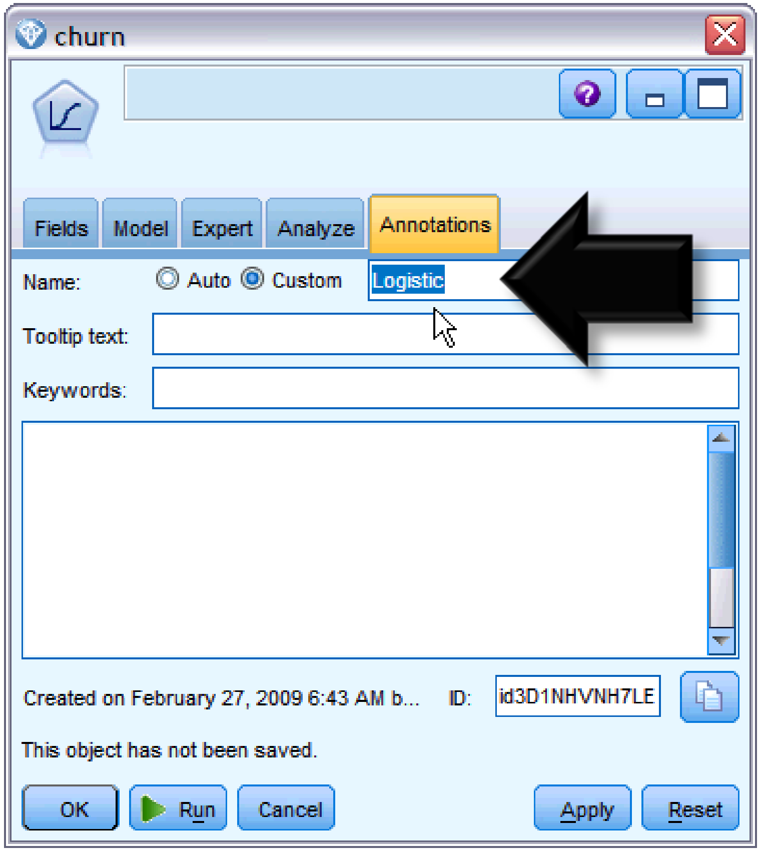


IBM® SPSS® Modeler

Version 14.2

# GUI: Comments, Custom Name Attribute, and SuperNodes Provide Communication, Security

- **Comments**
  - A sticky notes-like feature that allows users to document thoughts and share details directly on the canvas
- **Custom Name** in the Attributes Tab
  - Specifies the name assigned to the model that is created when the node is executed.
    - Auto - generates the model name automatically
    - Custom - allows you to specify a custom name for the model created
- **SuperNodes**
  - Creating a SuperNode "shrinks" the stream by encapsulating several nodes into one



Name: ○ Auto ● Custom Logistic

Tooltip text:

Keywords:

Created on February 27, 2009 6:43 AM b... ID: id3D1NHVNH7LE

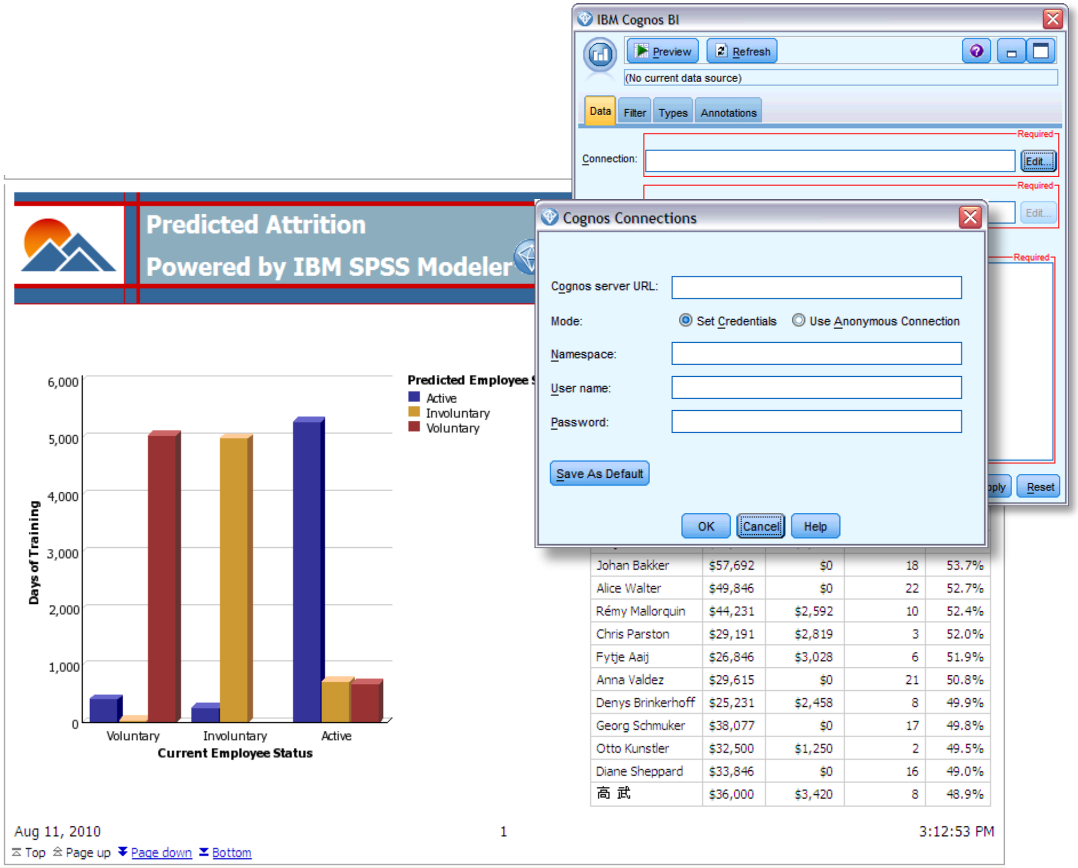
This object has not been saved.

# GUI: Other Interface Shortcuts that Speed the Development and Execution of SPSS Modeler Streams

- Build streams quickly by double-clicking
  - Double-clicking a node on the palette will add and connect it to the current stream
- Use key combinations to select all downstream nodes
  - Pressing Ctrl-Q and Ctrl-W will toggle the selection of <u>all</u> nodes downstream.
- Use shortcut keys to connect and disconnect nodes
  - When a node is selected in the canvas, pressing:
    - F2 will begin a connection
    - Tab will move to the desired node
    - Shift-spacebar will complete the connection
    - F3 will disconnect all inputs and outputs to the selected node
- Customize the Nodes Palette tab with your favorite nodes
  - From the Tools menu, choose Manage Palettes to open a dialog box for adding, removing, or moving the nodes shown on the Nodes Palette.

# Sources: Importing, Exporting to Cognos BI Enables a Distribution of Predictive Analytics based on “One Version of the Truth”

- **IBM Cognos BI Source Node**
  - Bring Cognos BI database data into your data mining session
  - Combine business intelligence features with the predictive analytics capabilities
  - Import relational, dimensionally-modeled relational and OLAP data
- **IBM Cognos BI Export Node**
  - Export data from a stream to Cognos BI (UTF-8 format)
  - Enables Cognos BI to make use of transformed or scored data from Modeler
    - Specifically, saved to a Cognos BI server and distributed to Cognos users



IBM Cognos BI

(No current data source)

Connection: Required

Cognos Connections

Cognos server URL:

Mode: ● Set Credentials ○ Use Anonymous Connection

Namespace:

User name:

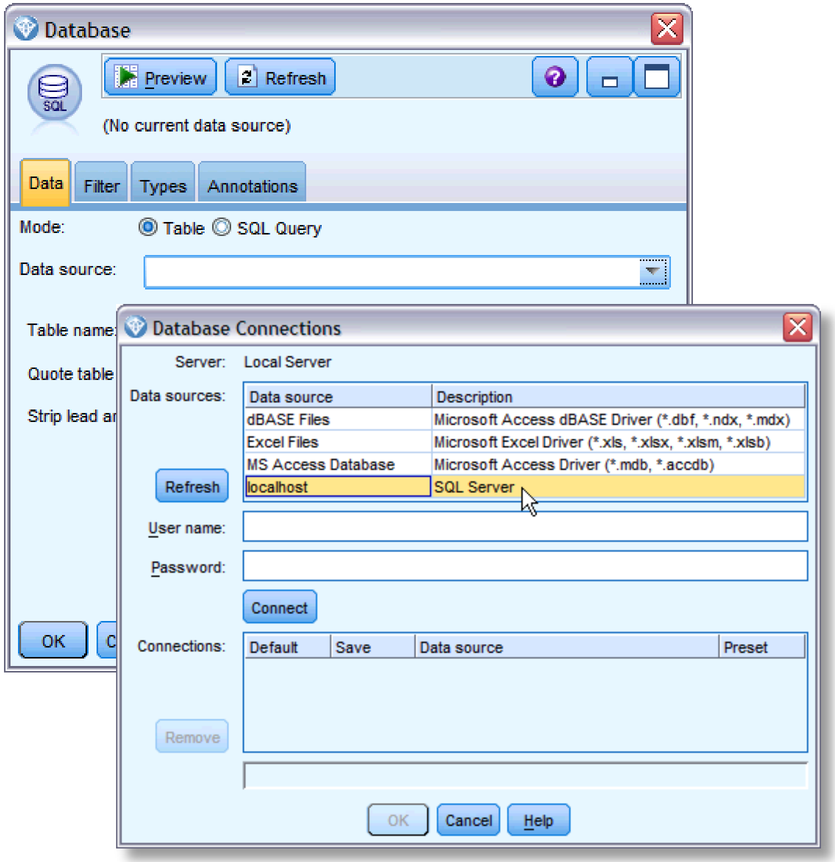
Password:

| | | | | |
|---|---|---|---|---|
| Johan Bakker | $57,692 | $0 | 18 | 53.7% |
| Alice Walter | $49,846 | $0 | 22 | 52.7% |
| Rémy Mallorquin | $44,231 | $2,592 | 10 | 52.4% |
| Chris Parston | $29,191 | $2,819 | 3 | 52.0% |
| Fytje Aaij | $26,846 | $3,028 | 6 | 51.9% |
| Anna Valdez | $29,615 | $0 | 21 | 50.8% |
| Denys Brinkerhoff | $25,231 | $2,458 | 8 | 49.9% |
| Georg Schmuker | $38,077 | $0 | 17 | 49.8% |
| Otto Kunstler | $32,500 | $1,250 | 2 | 49.5% |
| Diane Sheppard | $33,846 | $0 | 16 | 49.0% |
| 高 武 | $36,000 | $3,420 | 8 | 48.9% |

Aug 11, 2010 1 3:12:53 PM

# Sources: Save Time Connecting to a Preferred Database by Specifying a Default Database

- **Database Connection** - For database import and export users can specify a default connection and store details of different connections for display. A few highlights:
  - Data Sources
    - Lists the available data sources
  - Connections
    - Shows currently connected databases
  - Default
    - Optionally choose one connection as the default causing
    - This setting can be edited if desired.
  - Preset
    - Indicates (with a * character) whether preset values have been specified for the database connection



(No current data source)

Mode: ● Table ○ SQL Query

Data source:

Server: Local Server

Data sources:

| Data source | Description |
|---|---|
| dBASE Files | Microsoft Access dBASE Driver (*.dbf, *.ndx, *.mdx) |
| Excel Files | Microsoft Excel Driver (*.xls, *.xlsx, *.xlsm, *.xlsb) |
| MS Access Database | Microsoft Access Driver (*.mdb, *.accdb) |
| localhost | SQL Server |

User name:

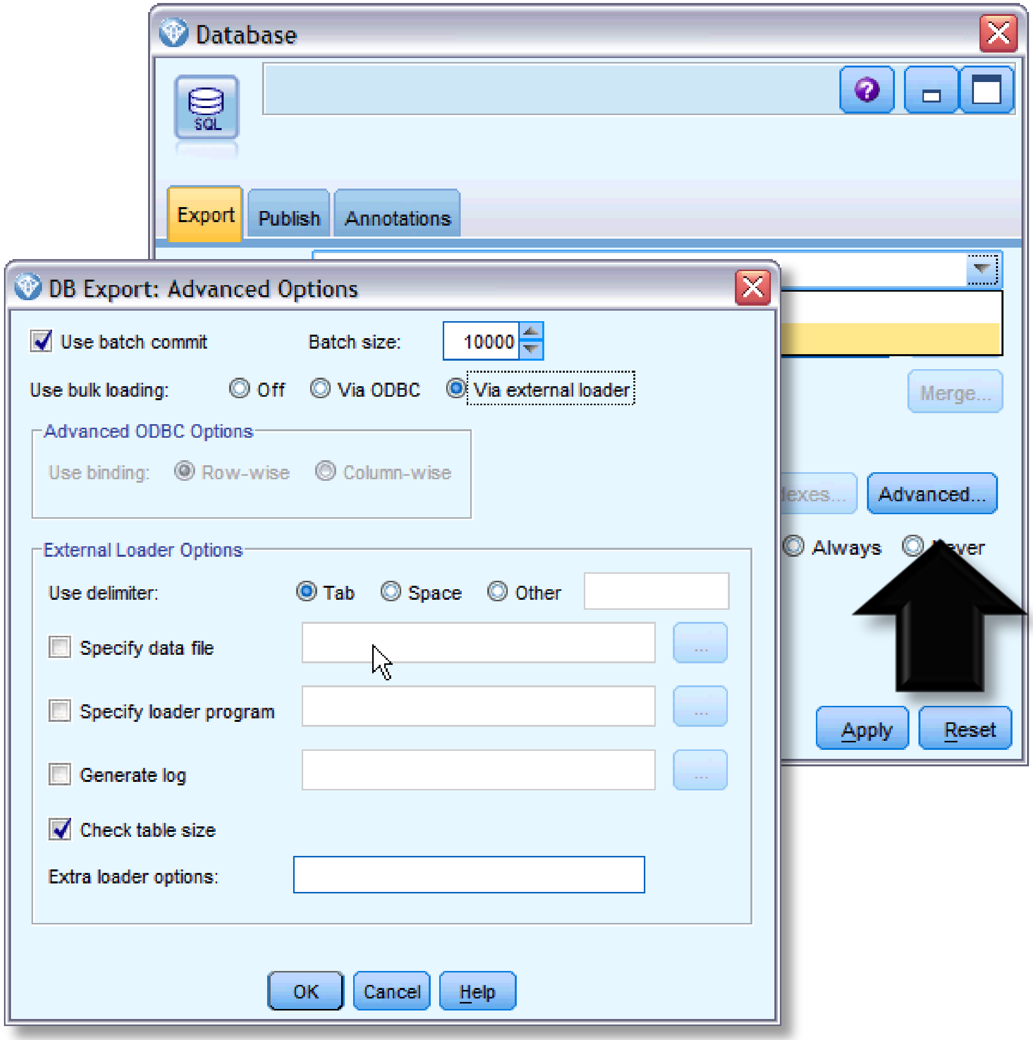
Password:

Connections:

| Default | Save | Data source | Preset |
|---|---|---|---|

# Sources: Options in Sub-Dialog Boxes also help Save Time when Connecting to a Databases

- **Database Export Node**
  - The “Schema” sub-dialog box helps improve efficiency when exporting to an InfoSphere Warehouse database
  - The “Advanced” sub-dialog box enables technical details for exporting results to a database



DB Export: Advanced Options

☑ Use batch commit — Batch size: 10000

Use bulk loading: ○ Off ○ Via ODBC ● Via external loader

Advanced ODBC Options

Use binding: ● Row-wise ○ Column-wise

External Loader Options

Use delimiter: ● Tab ○ Space ○ Other

☐ Specify data file

☐ Specify loader program

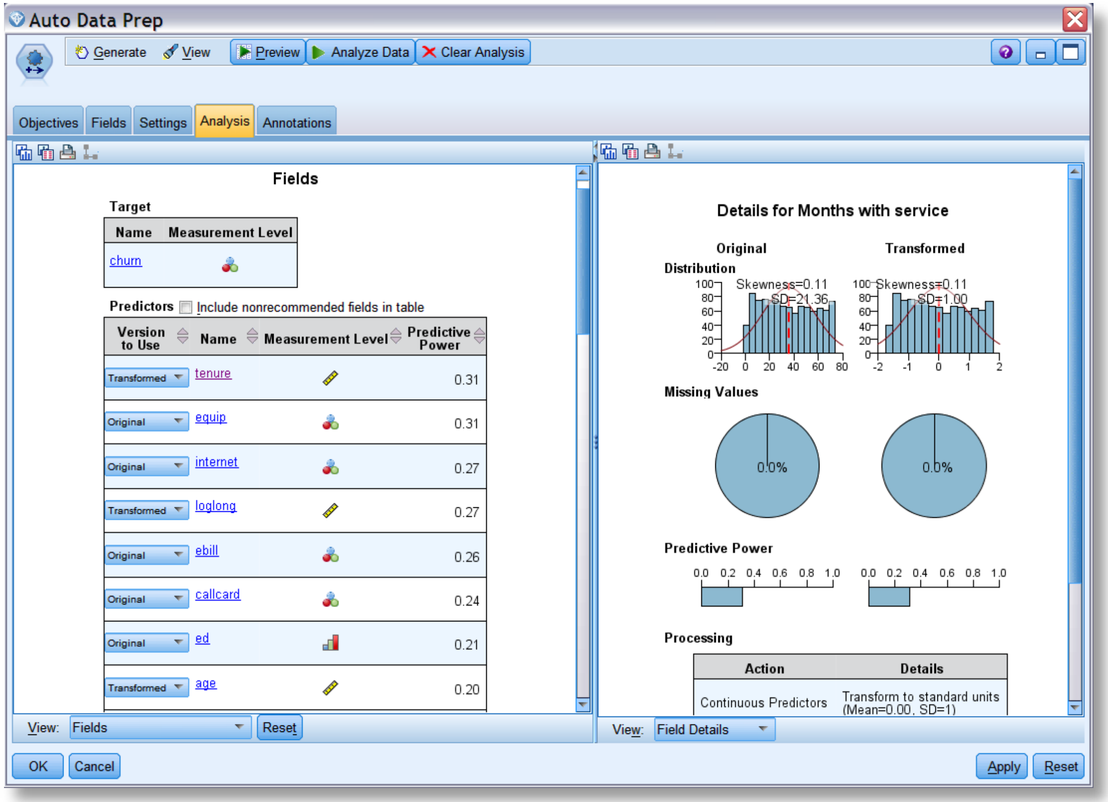
☐ Generate log

☑ Check table size

Extra loader options:

# Data Prep: Reduce Time Spent Preparing Data for Analysis with Automated Data Preparation

- **Automated Data Preparation (ADP) Node**
  - Analyzes data
  - Identifies possible fixes
  - Screens out fields that are problematic or not likely to be useful
  - Derives new attributes when appropriate
  - Improves performance through intelligent screening techniques



**Fields**

Target

| Name | Measurement Level |
|---|---|
| churn | ● |

Predictors □ Include nonrecommended fields in table

| Version to Use | Name | Measurement Level | Predictive Power |
|---|---|---|---|
| Transformed | tenure | 📏 | 0.31 |
| Original | equip | ● | 0.31 |
| Original | internet | ● | 0.27 |
| Transformed | loglong | 📏 | 0.27 |
| Original | ebill | ● | 0.26 |
| Original | callcard | ● | 0.24 |
| Original | ed | 📊 | 0.21 |
| Transformed | age | 📏 | 0.20 |

View: Fields

**Details for Months with service**

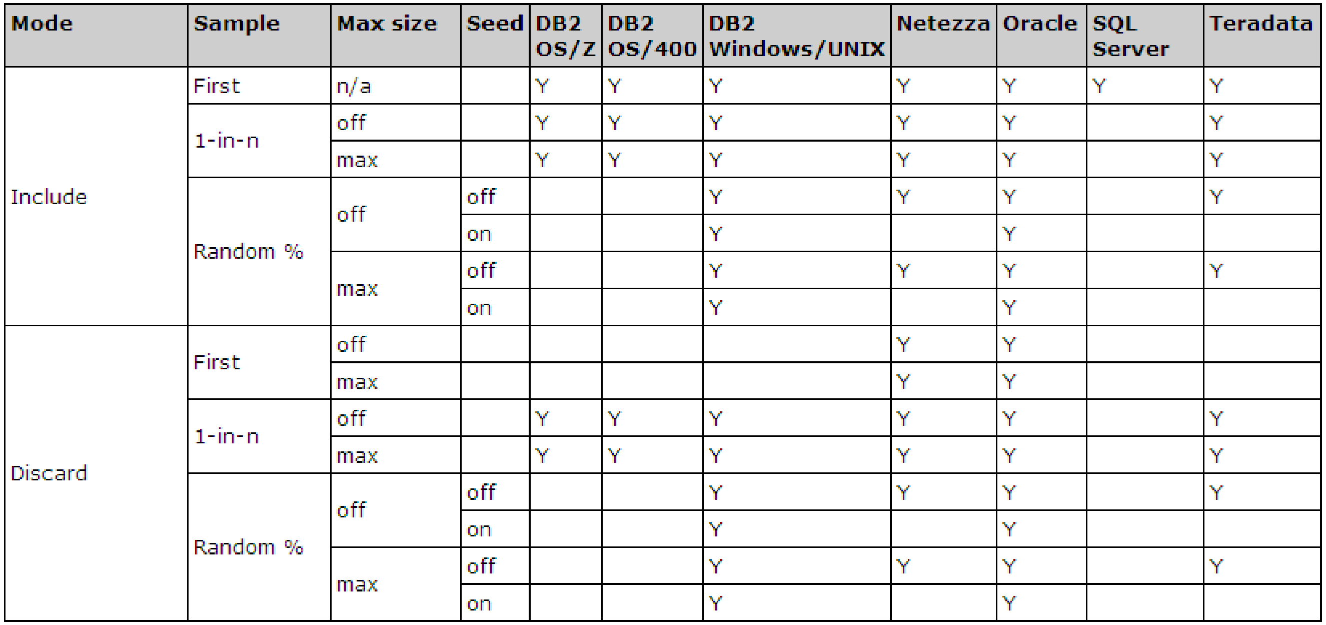
Processing

| Action | Details |
|---|---|
| Continuous Predictors | Transform to standard units (Mean=0.00, SD=1) |

View: Field Details

# Data Prep: Enhancements to the Sample Node include Data Mining that is more Efficient and Effective

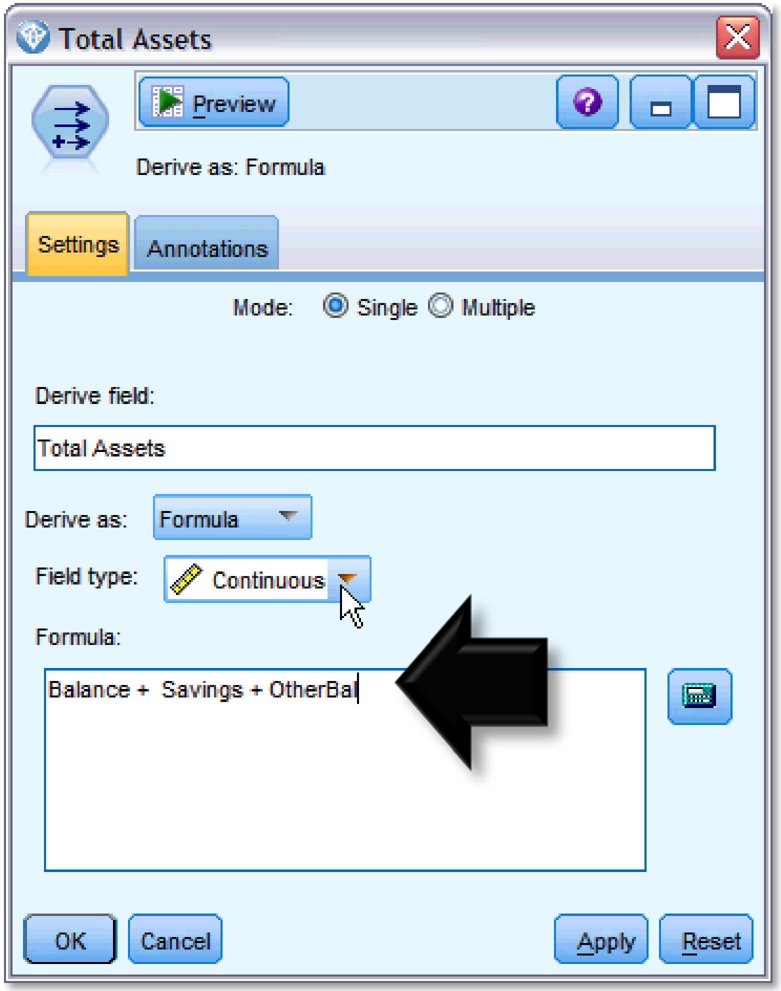
- **Sample Node**
  - Oracle or IBM DB2 database
    - Block-level sampling can be more efficient with random percentage sampling when performing in-database mining
  - Enhanced support for SQL generation (14.2)
    - Support for SQL generation in the Sample node when using Simple sampling has been enhanced for DB2, Netezza, and Teradata databases



| Mode | Sample | Max size | Seed | DB2 OS/Z | DB2 OS/400 | DB2 Windows/UNIX | Netezza | Oracle | SQL Server | Teradata |
|---|---|---|---|---|---|---|---|---|---|---|
| Include | First | n/a |  | Y | Y | Y | Y | Y | Y | Y |
|  | 1-in-n | off |  | Y | Y | Y | Y | Y |  | Y |
|  |  | max |  | Y | Y | Y | Y | Y |  | Y |
|  | Random % | off | off |  |  | Y | Y | Y |  | Y |
|  |  |  | on |  |  | Y |  | Y |  |  |
|  |  | max | off |  |  | Y | Y | Y |  | Y |
|  |  |  | on |  |  | Y |  | Y |  |  |
| Discard | First | off |  |  |  |  | Y | Y |  |  |
|  |  | max |  |  |  |  | Y | Y |  |  |
|  | 1-in-n | off |  | Y | Y | Y | Y | Y |  | Y |
|  |  | max |  | Y | Y | Y | Y | Y |  | Y |
|  | Random % | off | off |  |  | Y | Y | Y |  | Y |
|  |  |  | on |  |  | Y |  | Y |  |  |
|  |  | max | off |  |  | Y | Y | Y |  | Y |
|  |  |  | on |  |  | Y |  | Y |  |  |

# Data Prep: Leveraging CLEM Speeds the Data Mining Process

- Background on CLEM
  - Control Language for Expression Manipulation
  - Powerful language for analyzing and manipulating the data that flows along streams
  - Perform both simple and complex tasks
    - Simple: Deriving profit from cost and revenue data
    - Complex: Transforming web log data into a set of fields and records with usable information.
- New to Version 14.1
  - Within the CLEM language a number of date and time functions now additionally support the use of timestamp as an argument



Total Assets

Derive as: Formula

Mode: ● Single ○ Multiple

Derive field: Total Assets

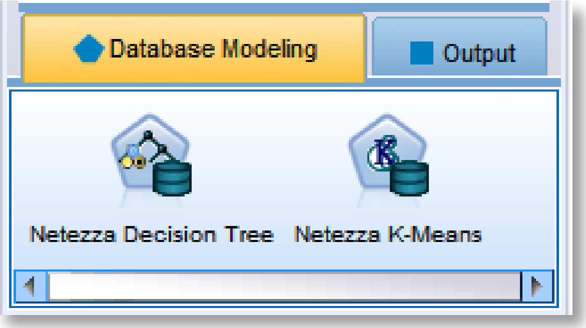
Derive as: Formula

Field type: Continuous

Formula: Balance + Savings + OtherBal

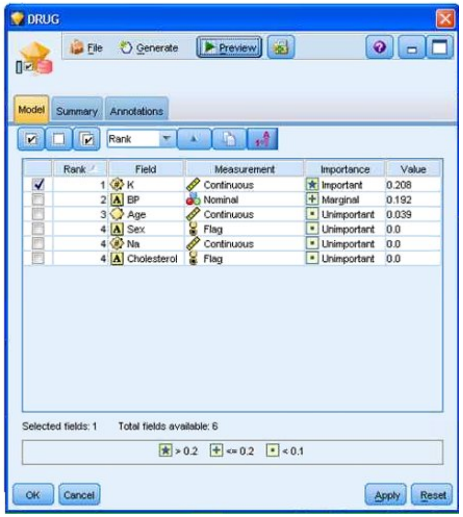
# Modeling: In-Database Mining Offers Improved Performance and Easy Deployment

- **In-Database Mining**
  - Server supports integration with data mining and modeling tools that are available from database vendors, including:
    - Oracle Data Miner
    - IBM DB2 InfoSphere Warehouse
    - Microsoft Analysis Services
    - IBM Netezza.
  - Build, score, and store models inside the database—all from within the Modeler
- Advantages of database-native algorithms include:
  - Improved performance
  - Easy deployment to / sharing with application that access the database



- **Microsoft Time Series**
- **Microsoft Sequence Clustering**
- **Oracle Attribute Importance**

Oracle AI

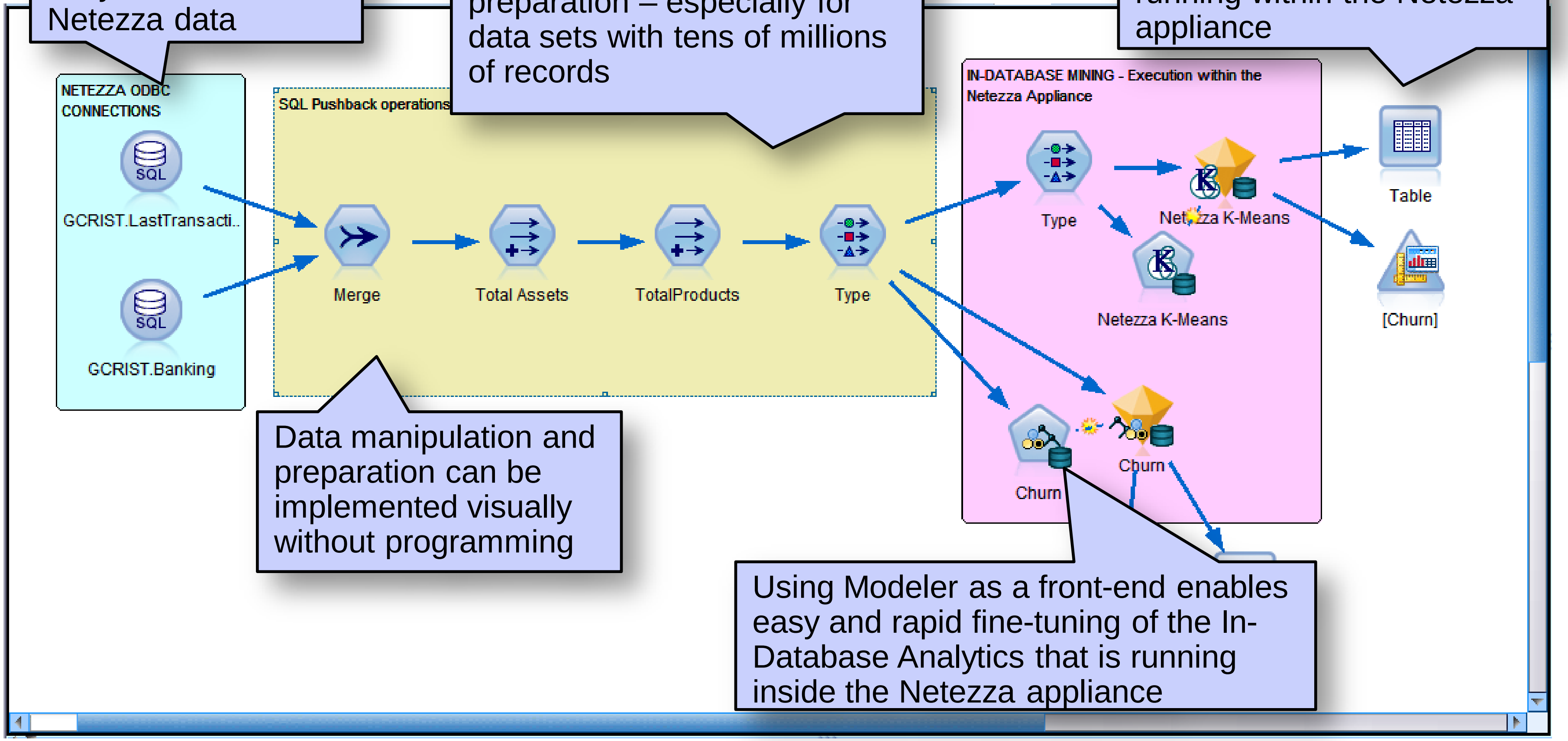


| | Rank | Field | Measurement | Importance | Value |
|---|---|---|---|---|---|
| ☑ | 1 | K | Continuous | Important | 0.208 |
| □ | 2 | BP | Nominal | Marginal | 0.192 |
| □ | 3 | Age | Continuous | Unimportant | 0.039 |
| □ | 4 | Sex | Flag | Unimportant | 0.0 |
| □ | 4 | Na | Continuous | Unimportant | 0.0 |
| □ | 4 | Cholesterol | Flag | Unimportant | 0.0 |

Selected fields: 1 Total fields available: 6

> 0.2 <= 0.2 < 0.1

# Modeling: Using SPSS Modeler as a Front-end is a Primary Benefit of Integration with a Database (Example: Netezza)



Using Modeler as a front-end enables easy connection to Netezza data

A visual front-end coupled with SQL pushback means rapid preparation – especially for data sets with tens of millions of records

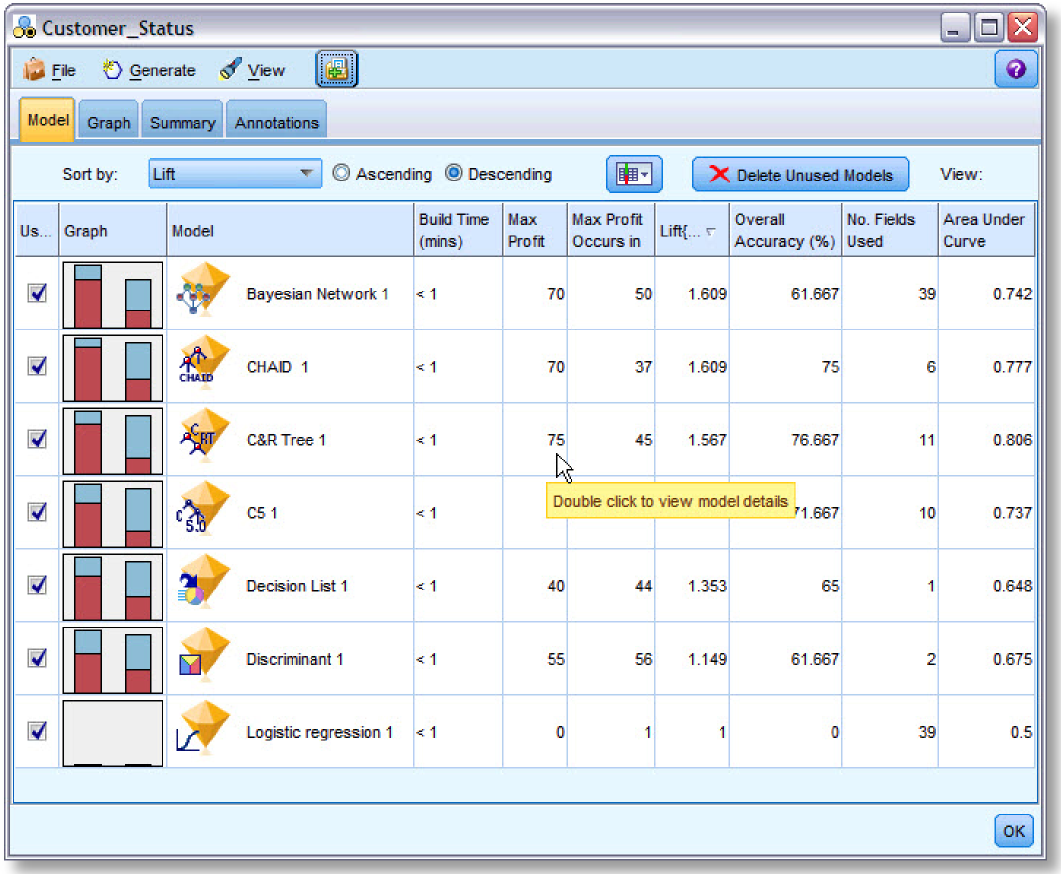
Modeler's various output nodes provide an easy way to interact with analytics running within the Netezza appliance

Data manipulation and preparation can be implemented visually without programming

Using Modeler as a front-end enables easy and rapid fine-tuning of the In-Database Analytics that is running inside the Netezza appliance

# Modeling: Automated Modeling Enables a Side-by-Side Comparison of Model Effectiveness

- **Auto Classifier, Auto Cluster, and Auto Numeric Nodes**
  - Standard a group of ensemble modeling nodes
  - Automate the building of a number of different models concurrently
  - Compare the results and choose the best model for data
  - CLEF provides the AutoModeling element to enable a model specified by the ModelBuilder element to be used by any of these ensemble nodes



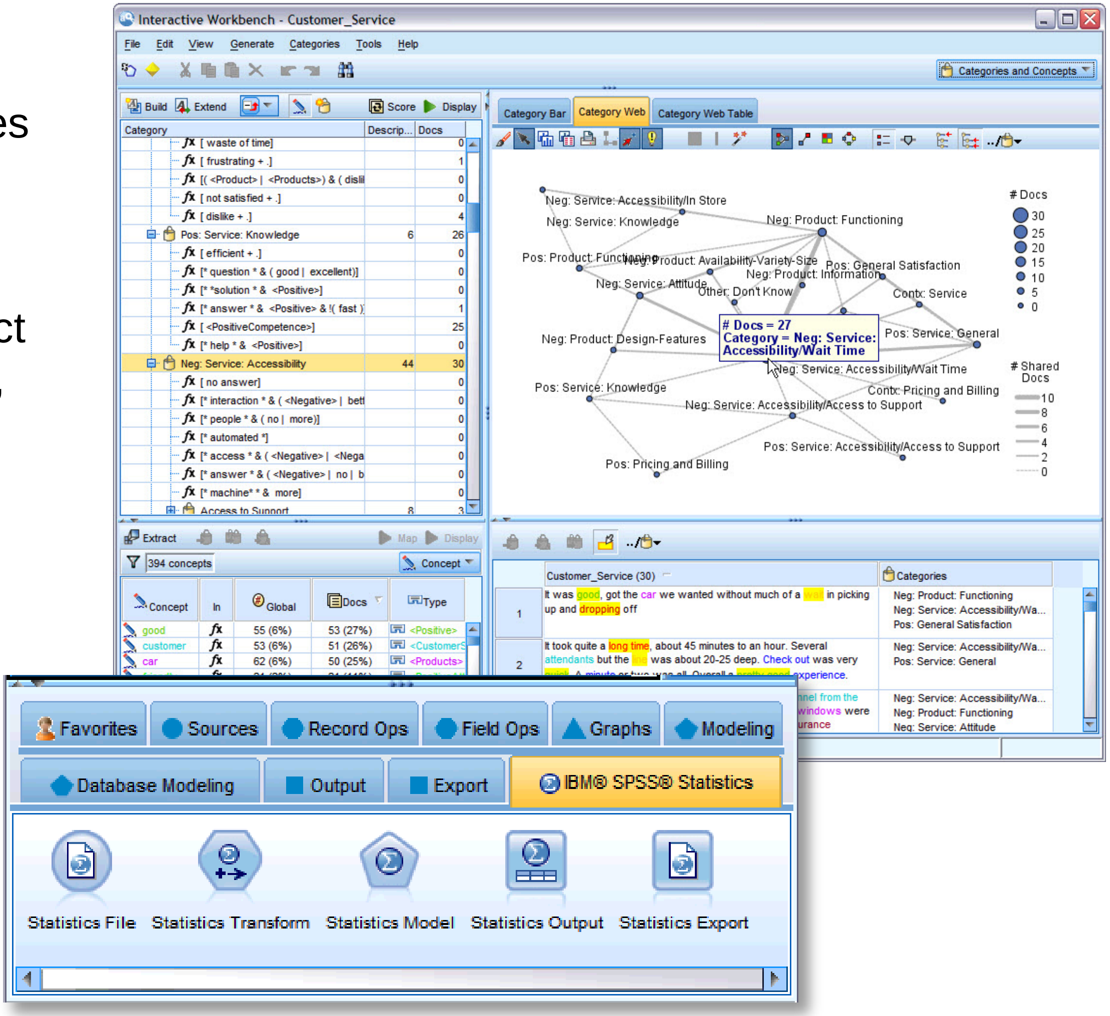
Sort by: Lift ○ Ascending ● Descending

| Us... | Graph | Model | Build Time (mins) | Max Profit | Max Profit Occurs in | Lift{... | Overall Accuracy (%) | No. Fields Used | Area Under Curve |
|---|---|---|---|---|---|---|---|---|---|
| ☑ | | Bayesian Network 1 | < 1 | 70 | 50 | 1.609 | 61.667 | 39 | 0.742 |
| ☑ | | CHAID 1 | < 1 | 70 | 37 | 1.609 | 75 | 6 | 0.777 |
| ☑ | | C&R Tree 1 | < 1 | 75 | 45 | 1.567 | 76.667 | 11 | 0.806 |
| ☑ | | C5 1 | < 1 | | | | 71.667 | 10 | 0.737 |
| ☑ | | Decision List 1 | < 1 | 40 | 44 | 1.353 | 65 | 1 | 0.648 |
| ☑ | | Discriminant 1 | < 1 | 55 | 56 | 1.149 | 61.667 | 2 | 0.675 |
| ☑ | | Logistic regression 1 | < 1 | 0 | 1 | 1 | 0 | 39 | 0.5 |

Double click to view model details

# Integration: Modeler Now Fully Integrated with IBM SPSS Statistics, Modeler Premium now includes Text Analytics

- **IBM SPSS Text Analytics**
  - Advanced linguistic technologies and Natural Language Processing (NLP)
  - Rapidly process a large variety of unstructured text data, extract and organize the key concepts, and group these concepts into categories
  - Extracted concepts and categories can be combined with existing structured data
- **IBM SPSS Statistics**
  - Fully integrated within the Modeler environment


Interactive Workbench - Customer_Service
File Edit View Generate Categories Tools Help
Categories and Concepts
Build Extend Score Display
Category Bar Category Web Category Web Table
Neg: Service: Accessibility/In Store
Neg: Service: Knowledge
Neg: Product: Functioning
Pos: Product: Functioning
Neg: Product: Availability-Variety-Size
Pos: General Satisfaction
Neg: Product: Information
Neg: Service: Attitude
Other: Don't Know
Contx: Service
# Docs = 27
Category = Neg: Service: Accessibility/Wait Time
Pos: Service: General
Neg: Product: Design-Features
Neg: Service: Accessibility/Wait Time
Pos: Service: Knowledge
Contx: Pricing and Billing
Neg: Service: Accessibility/Access to Support
Pos: Service: Accessibility/Access to Support
Pos: Pricing and Billing
# Docs
30
25
20
15
10
5
0
# Shared Docs
10
8
6
4
2
0
Extract Map Display
394 concepts
Concept
Favorites Sources Record Ops Field Ops Graphs Modeling
Database Modeling Output Export IBM® SPSS® Statistics
Statistics File
Statistics Transform
Statistics Model
Statistics Output
Statistics Export

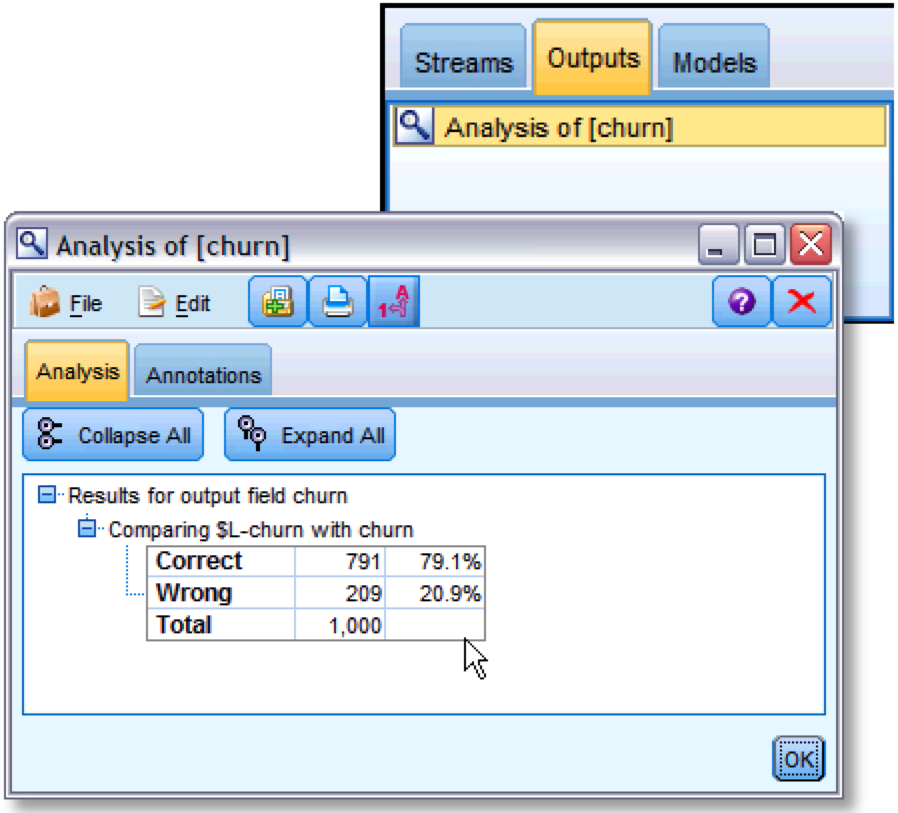
| Category | Descrip... | Docs |
|---|---|---|
| [ waste of time] | | 0 |
| [ frustrating + .] | | 1 |
| [( <Product> \| <Products>) & ( disli | | 0 |
| [ not satisfied + .] | | 0 |
| [ dislike + .] | | 4 |
| Pos: Service: Knowledge | 6 | 26 |
| [ efficient + .] | | 0 |
| [* question * & ( good \| excellent)] | | 0 |
| [* *solution * & <Positive>] | | 0 |
| [* answer * & <Positive> & !( fast ) | | 1 |
| [ <PositiveCompetence>] | | 25 |
| [* help * & <Positive>] | | 0 |
| Neg: Service: Accessibility | 44 | 30 |
| [ no answer] | | 0 |
| [* interaction * & ( <Negative> \| bett | | 0 |
| [* people * & ( no \| more)] | | 0 |
| [* automated *] | | 0 |
| [* access * & ( <Negative> \| <Nega | | 0 |
| [* answer * & ( <Negative> \| no \| b | | 0 |
| [* machine* * & more] | | 0 |
| Access to Support | 8 | 3 |

| Concept | In | Global | Docs | Type |
|---|---|---|---|---|
| good | fx | 55 (6%) | 53 (27%) | <Positive> |
| customer | fx | 53 (6%) | 51 (26%) | <CustomerS |
| car | fx | 62 (6%) | 50 (25%) | <Products> |

| | Customer_Service (30) | Categories |
|---|---|---|
| 1 | It was good, got the car we wanted without much of a wait in picking up and dropping off | Neg: Product: Functioning<br>Neg: Service: Accessibility/Wa...<br>Pos: General Satisfaction |
| 2 | It took quite a long time, about 45 minutes to an hour. Several attendants but the line was about 20-25 deep. Check out was very quick. A minute or two was all. Overall a pretty good experience. | Neg: Service: Accessibility/Wa...<br>Pos: Service: General |
| | nel from the windows were urance | Neg: Service: Accessibility/Wa...<br>Neg: Product: Functioning<br>Neg: Service: Attitude |

# Output: Analysis Node can be used to Compare Models, Results can be Written to a File

- **Analysis Node**
  - Evaluate the ability of a model to generate accurate predictions.
  - Perform various comparisons between predicted values and actual values
  - Tip: Analysis nodes can be used to compare predictive models to other predictive models
  - Upon execution a summary of the analysis results automatically added to the Analysis section on the Summary tab
  - Detailed analysis results appear on the Outputs tab of the manager window or can be written directly to a file



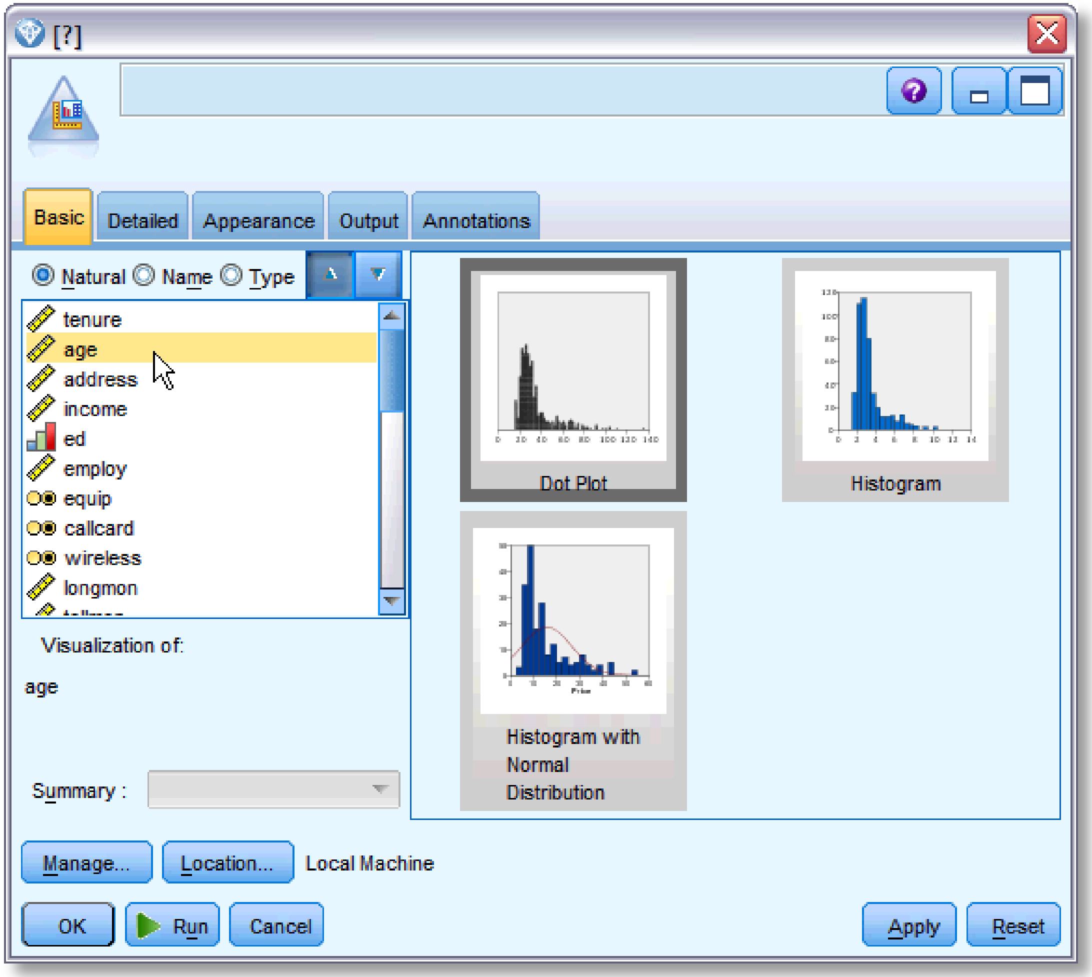
Results for output field churn

Comparing $L-churn with churn

|  |  |  |
|---|---|---|
| Correct | 791 | 79.1% |
| Wrong | 209 | 20.9% |
| Total | 1,000 |  |

# Output: The Graphboard Node Presents a Variety of Different Visualizations based on Data Types

- **Graphboard Node**
  - Choose from many different graphs outputs in a single node
  - Node presents you with a choice of graph types that work for your data
  - Automatically filters out any graph types that would not work with the field choices



Visualization of:

age

Summary :

Local Machine

# Recap: Agenda, Goals for the Session, and Next Steps

- Today’s Session included General Buckets Addressing:
  - Naming
  - GUI – the User Interface
  - Sources – Accessing Data
  - Data Preparation
  - Modeling
  - Integration
  - Output
- The Goal for Today’s Session was to
  - Take away at least one tip, trick, or fact about what’s new that will improve the efficiency and/or effectiveness of your data mining / predictive analytics efforts
- Next Steps
  - For More Information Call (800) 543-2185