IBM Analytics

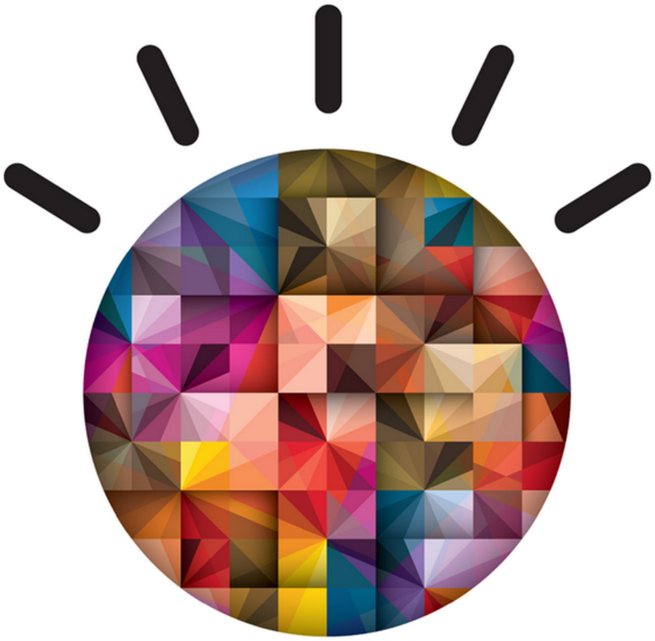
Commercial Workshop Guide

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

THE INFORMATION IN THIS DOCUMENT IS PROVIDED “AS IS” WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT. IBM products are warranted according to the terms and conditions of the agreements under which they are provided.

Table of Contents

Introduction



IBM SPSS Modeler is a comprehensive predictive analytics platform, designed to bring **predictive intelligence** to everyday business problems, enabling front-line employees to make more **effective decisions** and **improve outcomes.**

This enables organizations to improve business processes and help people or systems **consistently make the right decisions** by delivering recommended actions at the point of impact. The result is a **rapid return on investment** (ROI) and the ability to **proactively** and **repeatedly reduce costs** and **increase productivity**.

This hands-on Predictive Analytics Workshop is an instructor led session using IBM's data mining and predictive modeling software and is designed for those who are familiar with predictive analytics, as well as for beginners. Through this workshop you will experience firsthand how IBM SPSS Modeler works and how easy it is to implement predictive analytics.

Exercise 1: Predictive in 20 Minutes

Use Case

Goal: Identify who has responded to a marketing campaign

Approach:

- Use a data extract from a CRM
- Prepare data for modeling
- Define which fields to use
- Choose the modeling technique
- Automatically generate a model to identify who has responded
- Review results

Why?

- To save marketing cost and increase marketing response, identify those likely to respond and focus marketing efforts on those prospects.

Customer Reference

A US supermarket chain uses behavioral analytics to draw customers in, boosting response rate by 35% and lowering costs 25%.

Business challenge: This supermarket chain in the US was blanketing its regional markets with direct mailers to advertise new products, showcase low prices and deliver coupons, regardless of their relevance to the recipient. By understanding customer behavior and running more targeted campaigns, the organization hoped to increase response rates and reduce costs.

The transformation: The retailer is using a powerful analytics solution that changes its approach to decision making, creating a customer experience that wins out over new competitors and online shopping. Analyzing point-of-sale (POS) and demographics data, the solution allows the supermarket to segment its customer population, model future purchasing behaviors and execute more targeted marketing campaigns. The solution sets the stage for data-driven decisions in the supply chain, security and risk management, and other areas.

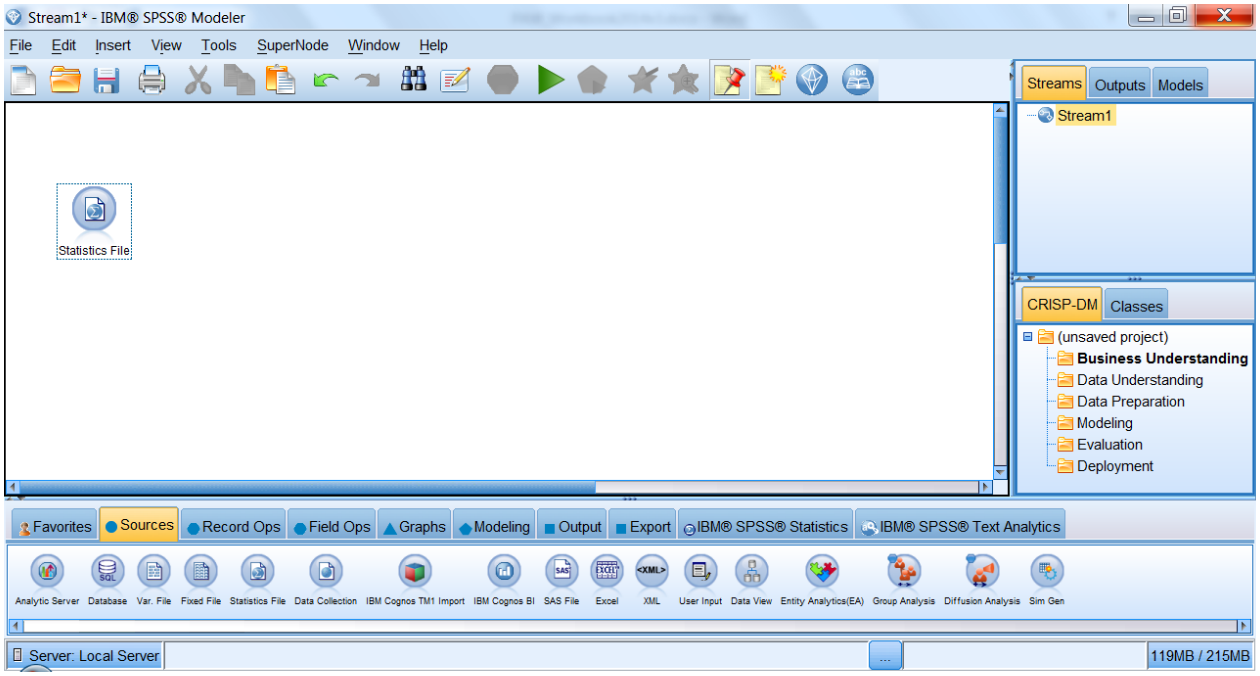
IBM's implemented solution resulted in:
35% increase in mailer response rates with more relevant content
25% lower costs by replacing mass marketing with targeted campaigns
99% faster reports, with the ability to analyze data in minutes instead of weeks

"The ability to analyze and predict customer behavior has become a necessity for retailers who rely on customer loyalty to stay ahead."

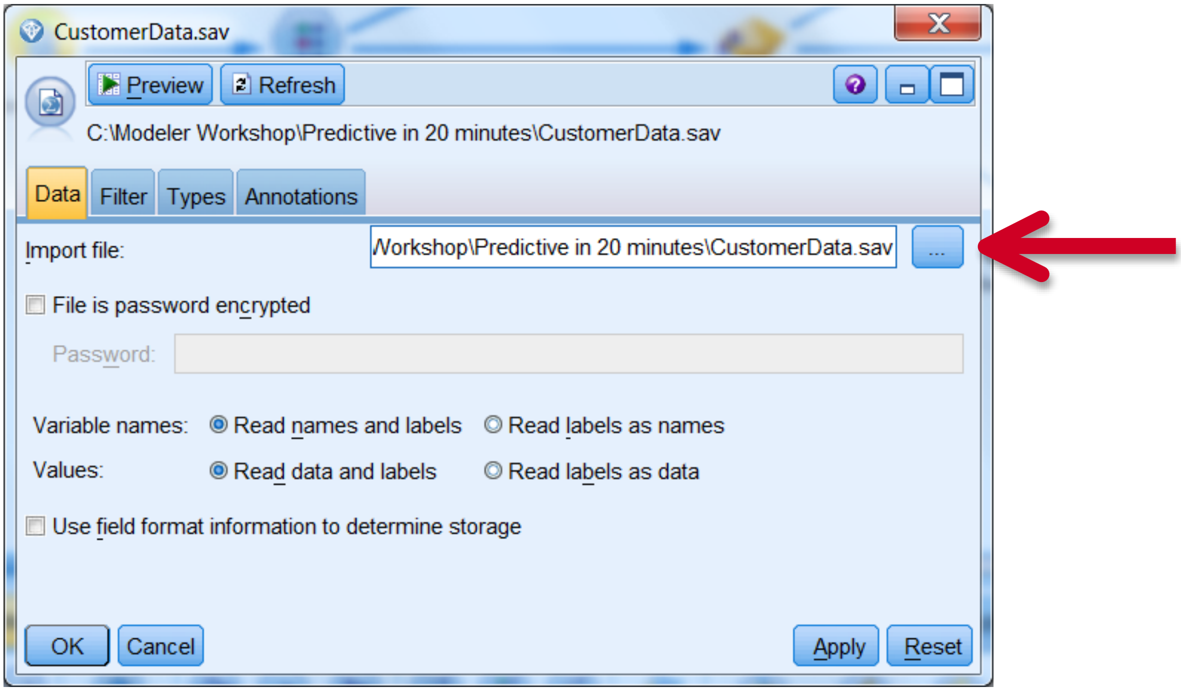
Predictive in 20 Minutes

1. Start Modeler if it's not already open.

2. From the Sources palette, double-click on the Statistics File node to add it to the canvas.


Stream1* - IBM® SPSS® Modeler
File Edit Insert View Tools SuperNode Window Help
Statistics File
Streams Outputs Models
Stream1
CRISP-DM Classes
(unsaved project)
Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Favorites Sources Record Ops Field Ops Graphs Modeling Output Export IBM® SPSS® Statistics IBM® SPSS® Text Analytics
Analytic Server Database Var. File Fixed File Statistics File Data Collection IBM Cognos TM1 Import IBM Cognos BI SAS File Excel XML User Input Data View Entity Analytics(EA) Group Analysis Diffusion Analysis Sim Gen
Server: Local Server
119MB / 215MB

3. Double-click the Statistics File node to open a dialog box. Use the data tab to import the Customer_Data.sav file from: C:\Modeler Workshop\Predictive in 20 Minutes\Customer_Data.sav.



C:\Modeler Workshop\Predictive in 20 minutes\CustomerData.sav

Import file: Workshop\Predictive in 20 minutes\CustomerData.sav ...

□ File is password encrypted

Password:

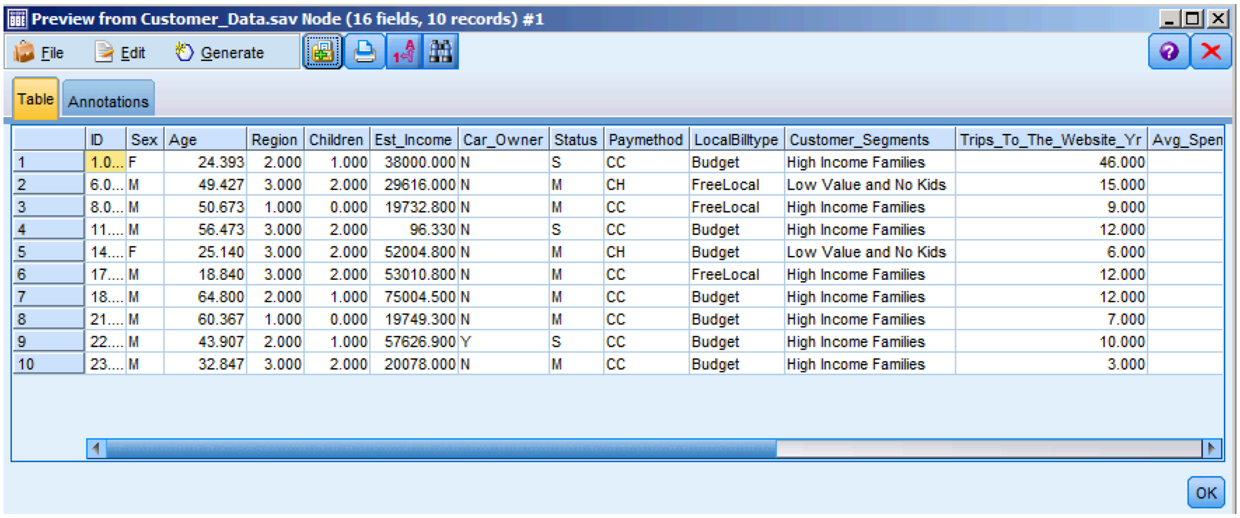
Variable names: ● Read names and labels ○ Read labels as names

Values: ● Read data and labels ○ Read labels as data

□ Use field format information to determine storage

© Copyright IBM Corporation 2016

4. Click on the Preview button at the top of the dialog box to see the first 10 records in the file, an extraction of data from a retail company's CRM system. It includes historical data related to their customers' demographics, purchasing behavior, segment, and marketing campaign response.

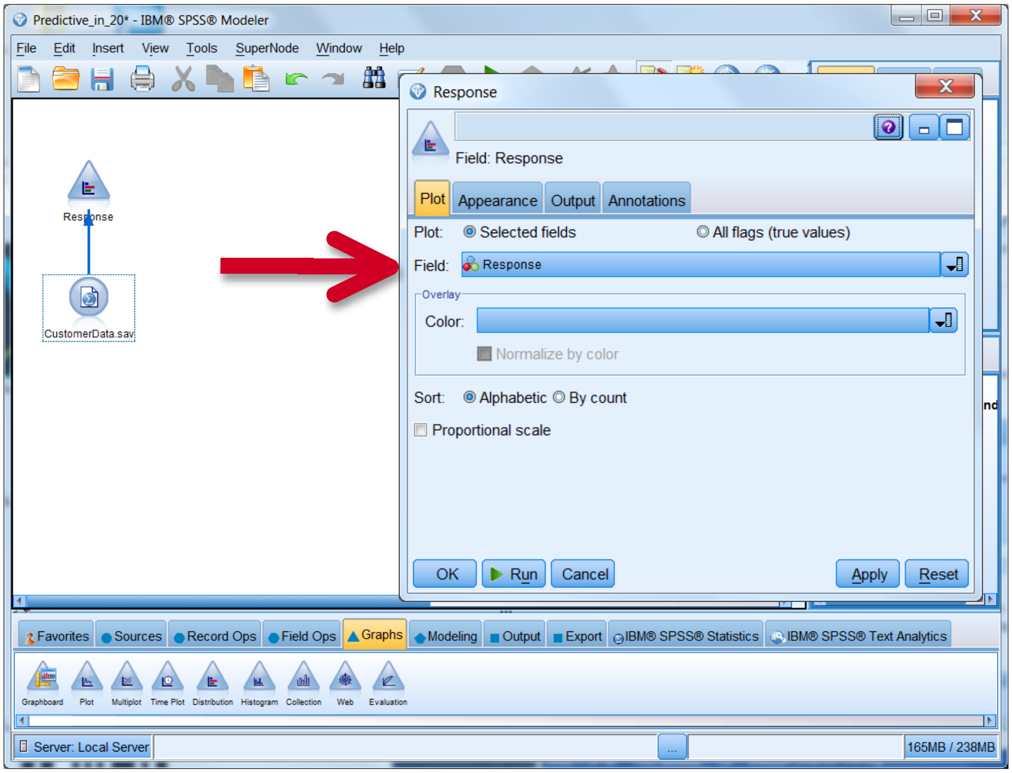

Preview from Customer_Data.sav Node (16 fields, 10 records) #1
File
Edit
Generate
Table
Annotations
OK

	ID	Sex	Age	Region	Children	Est_Income	Car_Owner	Status	Paymethod	LocalBilltype	Customer_Segments	Trips_To_The_Website_Yr	Avg_Spen
1	1.0...	F	24.393	2.000	1.000	38000.000	N	S	CC	Budget	High Income Families	46.000	
2	6.0...	M	49.427	3.000	2.000	29616.000	N	M	CH	FreeLocal	Low Value and No Kids	15.000	
3	8.0...	M	50.673	1.000	0.000	19732.800	N	M	CC	FreeLocal	High Income Families	9.000	
4	11....	M	56.473	3.000	2.000	96.330	N	S	CC	Budget	High Income Families	12.000	
5	14....	F	25.140	3.000	2.000	52004.800	N	M	CH	Budget	Low Value and No Kids	6.000	
6	17....	M	18.840	3.000	2.000	53010.800	N	M	CC	FreeLocal	High Income Families	12.000	
7	18....	M	64.800	2.000	1.000	75004.500	N	M	CC	Budget	High Income Families	12.000	
8	21....	M	60.367	1.000	0.000	19749.300	N	M	CC	Budget	High Income Families	7.000	
9	22....	M	43.907	2.000	1.000	57626.900	Y	S	CC	Budget	High Income Families	10.000	
10	23....	M	32.847	3.000	2.000	20078.000	N	M	CC	Budget	High Income Families	3.000	

5. After reviewing the Preview, or any subsequent output, click on the red X to close.

6. From the Graphs palette, add a distribution node to the canvas and connect it to the data source using any of the following methods:
   - Double-click on the node in the palette to automatically add it to the stream and join it to the selected node.
   - Drag and drop the node from out of the palette and on to the canvas. Select the first node, right-click and select Connect from the context menu, and then left-click on the second node to connect it.
   - Click and hold the middle mouse button on the first node, move the cursor to the second node and release when the cursor is on top of the second node.

Double-click to edit the Distribution node, choose Response from the Field drop down menu, and select Run.



Response

Field: Response

Plot: ● Selected fields ○ All flags (true values)

Field: Response

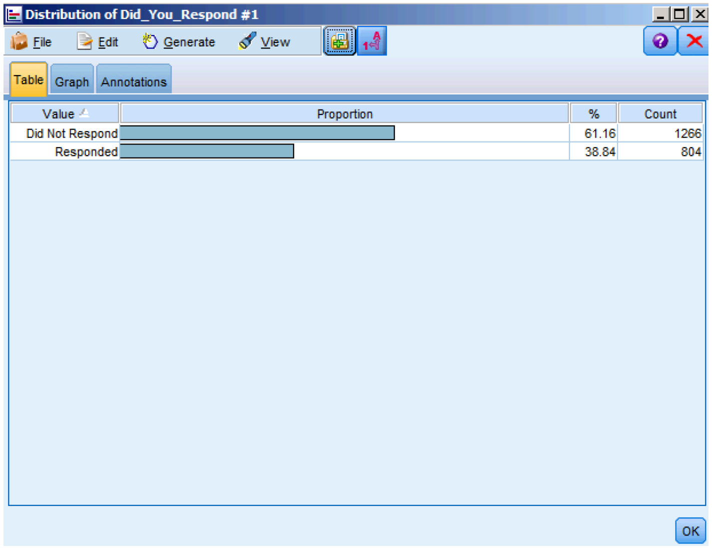
Color:

□ Normalize by color

Sort: ● Alphabetic ○ By count

□ Proportional scale

7. The resulting graph shows that of the 2070 customers in this dataset, only 38.84%, or 804 customers, responded to the campaign. The remaining 61.16%, or 1266 customers, did not respond. Our goal, then, is to build a model to understand the relationships within the data that lent themselves to customer response.

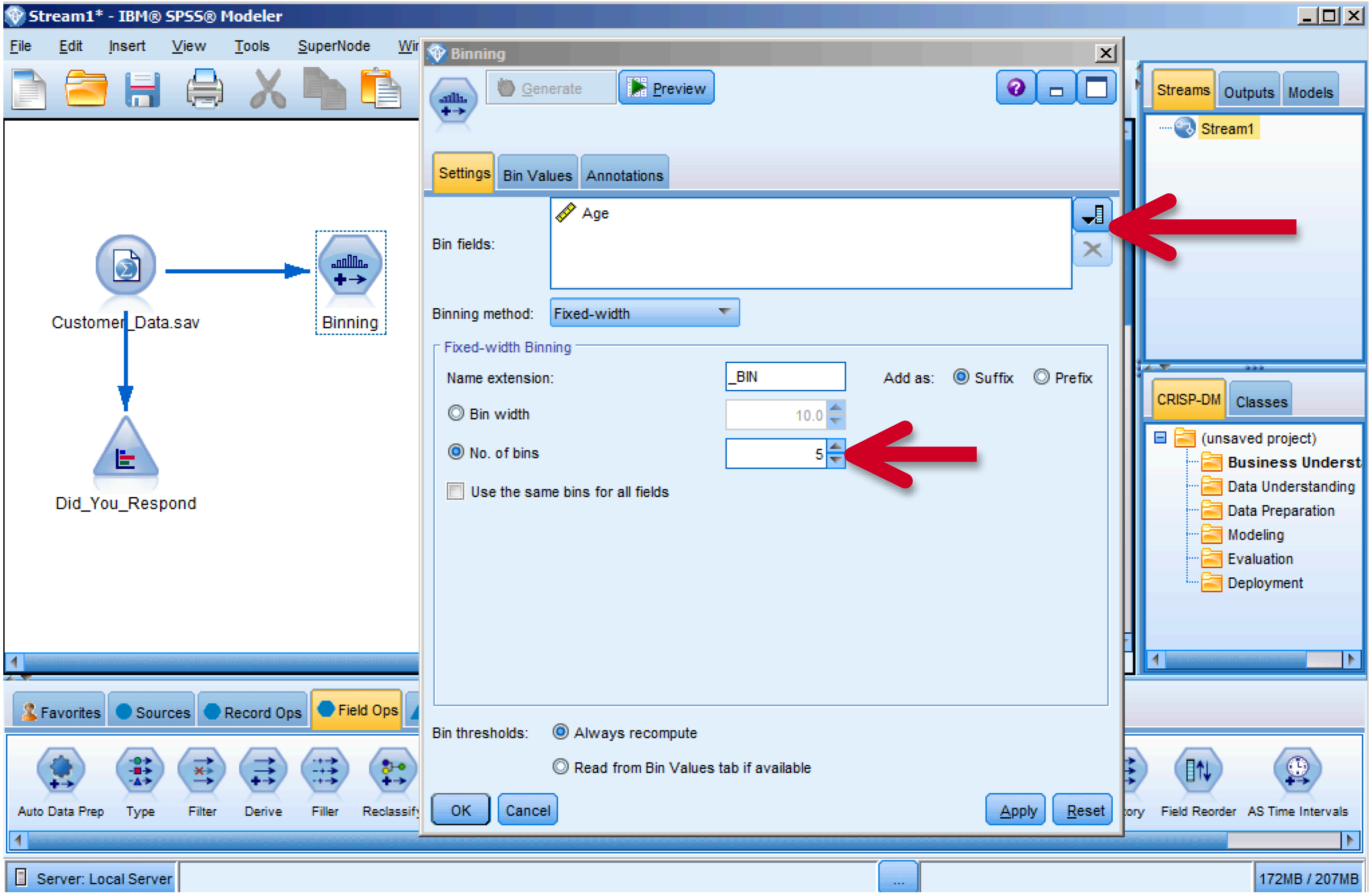


Distribution of Did_You_Respond #1

Value	Proportion	%	Count
Did Not Respond		61.16	1266
Responded		38.84	804

8. From the Field Ops palette, add a Binning node to the canvas and connect it to the data source. The Binning node allows you to automatically generate bins (categories) using several techniques. In this case, we will be creating categories from the continuous variable Age.

Double-click on the Binning node to edit the settings. Using the select field icon, select Age as the Bin field. Leaving the Binning method at fixed-width, select 5 as the No of bins to create, then click OK. The Bin Values tab (not shown) allows you to see the lower and upper cut points. By selecting Preview (not shown), you can see the appended field Age_Bin, which shows 5 possible categories.



Bin fields: Age

Binning method: Fixed-width

Fixed-width Binning

Name extension: _BIN   Add as: ● Suffix ○ Prefix

○ Bin width 10.0

● No. of bins 5

□ Use the same bins for all fields

Bin thresholds: ● Always recompute

○ Read from Bin Values tab if available

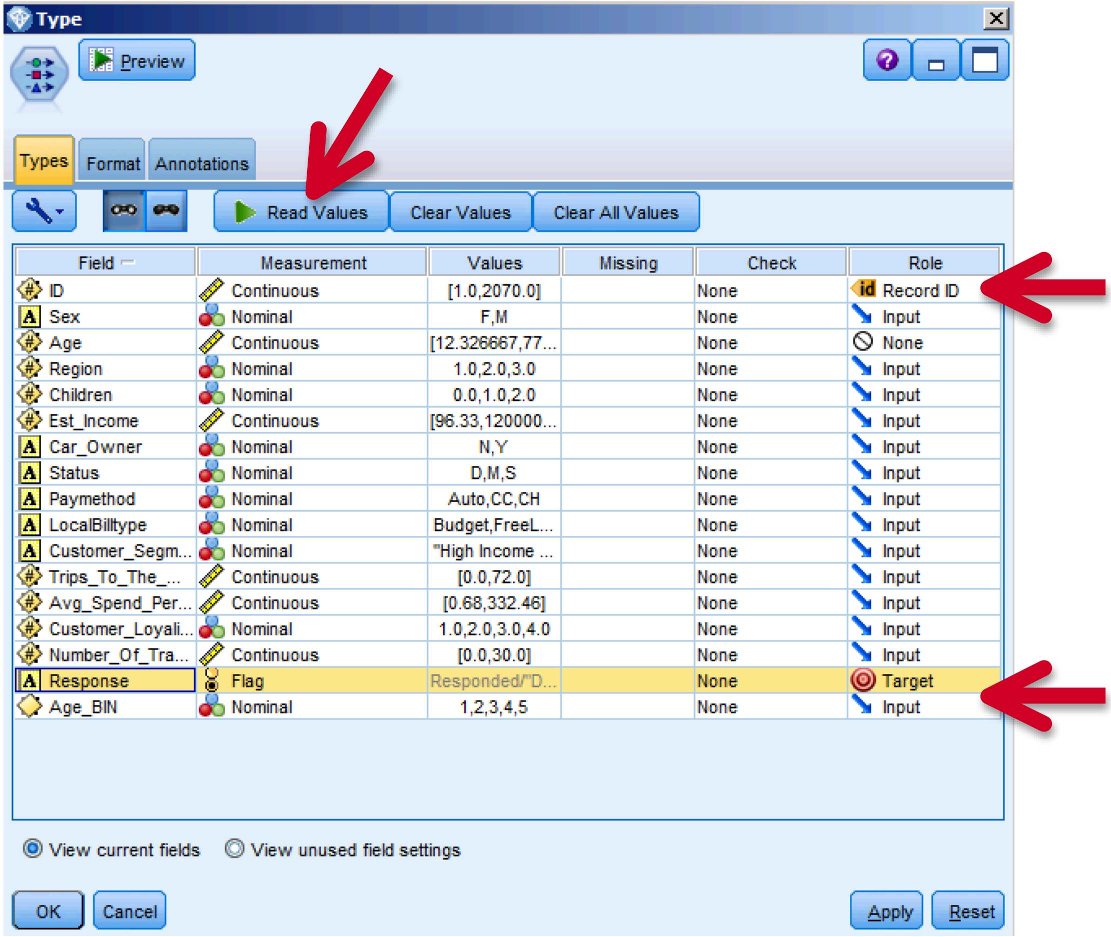
9. From the Field Ops palette, add a Type node to the canvas and connect it to the data source.

Double-click on the Type node and click the Read Values button to scan the data as well as to display and update the range of values (instantiate).

Using the drop-down box under Role, modify the following Fields:

- ID = Record ID
- Age = None
- Response = Target

The Measurement of our Target should be set to Flag, which reflects two potential Values: Responded or Did not Respond. The remaining, including our new Age_Bin, will remain as Inputs in our analysis.

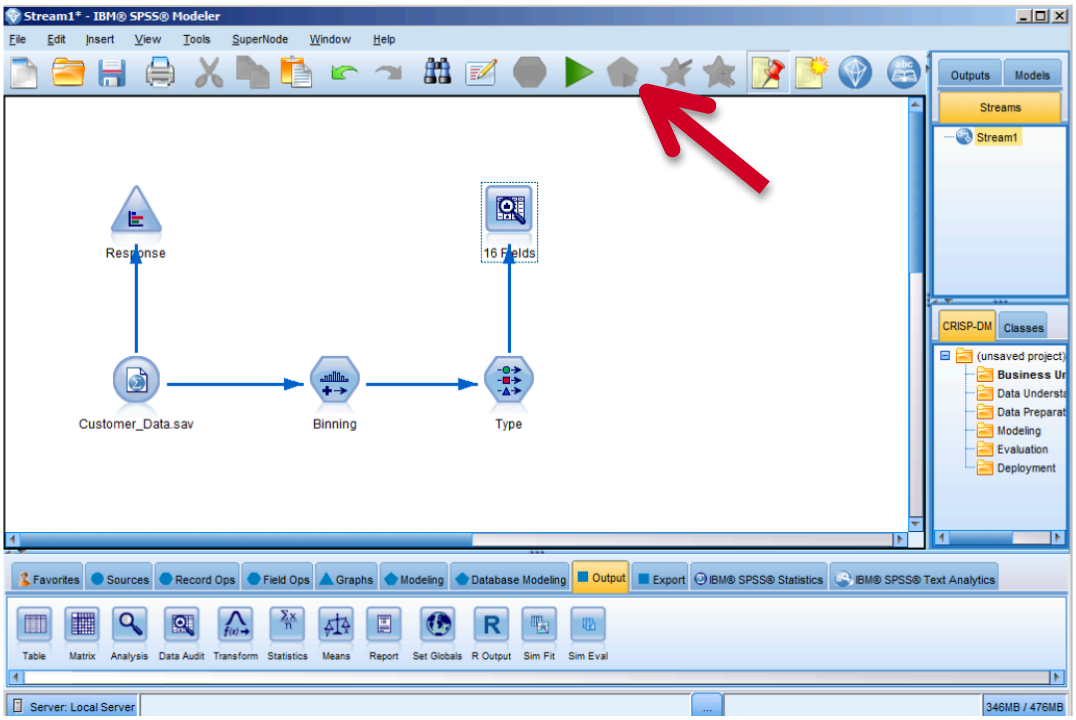


Field	Measurement	Values	Missing	Check	Role
ID	Continuous	[1.0,2070.0]		None	Record ID
Sex	Nominal	F,M		None	Input
Age	Continuous	[12.326667,77...		None	None
Region	Nominal	1.0,2.0,3.0		None	Input
Children	Nominal	0.0,1.0,2.0		None	Input
Est_Income	Continuous	[96.33,120000...		None	Input
Car_Owner	Nominal	N,Y		None	Input
Status	Nominal	D,M,S		None	Input
Paymethod	Nominal	Auto,CC,CH		None	Input
LocalBilltype	Nominal	Budget,FreeL...		None	Input
Customer_Segm...	Nominal	"High Income ...		None	Input
Trips_To_The_...	Continuous	[0.0,72.0]		None	Input
Avg_Spend_Per...	Continuous	[0.68,332.46]		None	Input
Customer_Loyali...	Nominal	1.0,2.0,3.0,4.0		None	Input
Number_Of_Tra...	Continuous	[0.0,30.0]		None	Input
Response	Flag	Responded/"D...		None	Target
Age_BIN	Nominal	1,2,3,4,5		None	Input

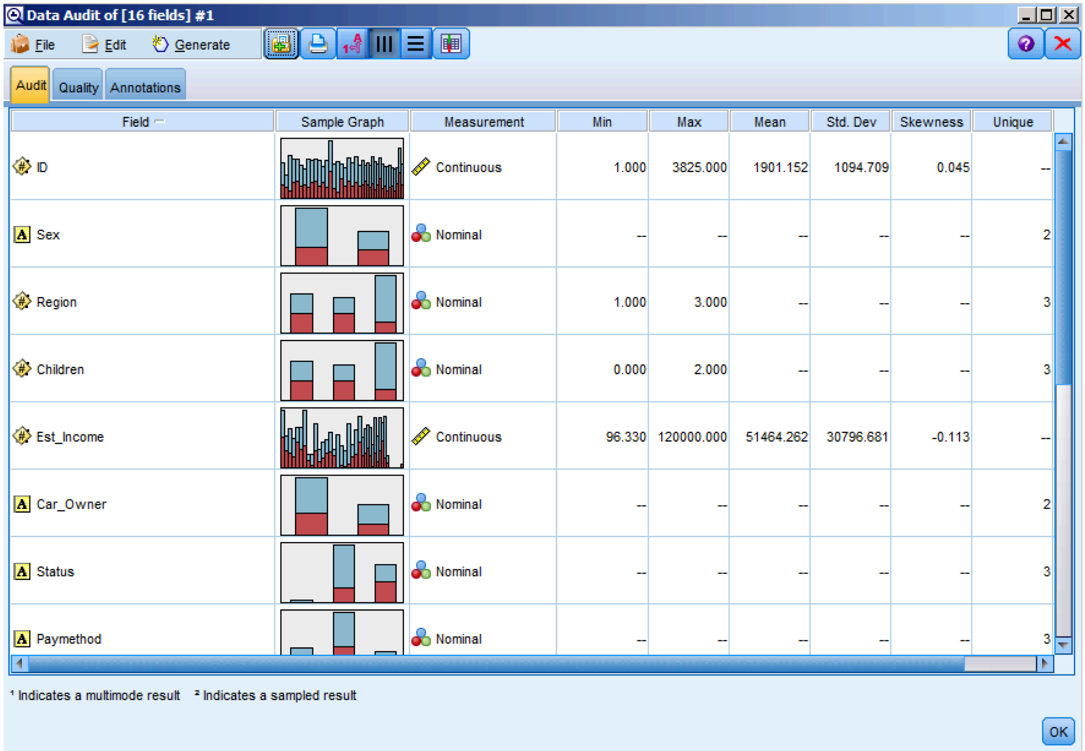
● View current fields ○ View unused field settings

10. From the Output palette, double-click on the Data Audit node and connect it to the Type node.

Right-click to select Run or chose the Run Selection button on the toolbar.



11. In the Audit tab of the resulting output, thumbnail graphs, storage icons, and summary statistics for all fields can be found. Double-clicking on any of the graphs will provide a more detailed outlook of the Field. In the Quality tab (not shown), information about outliers, extremes and missing values are shown.

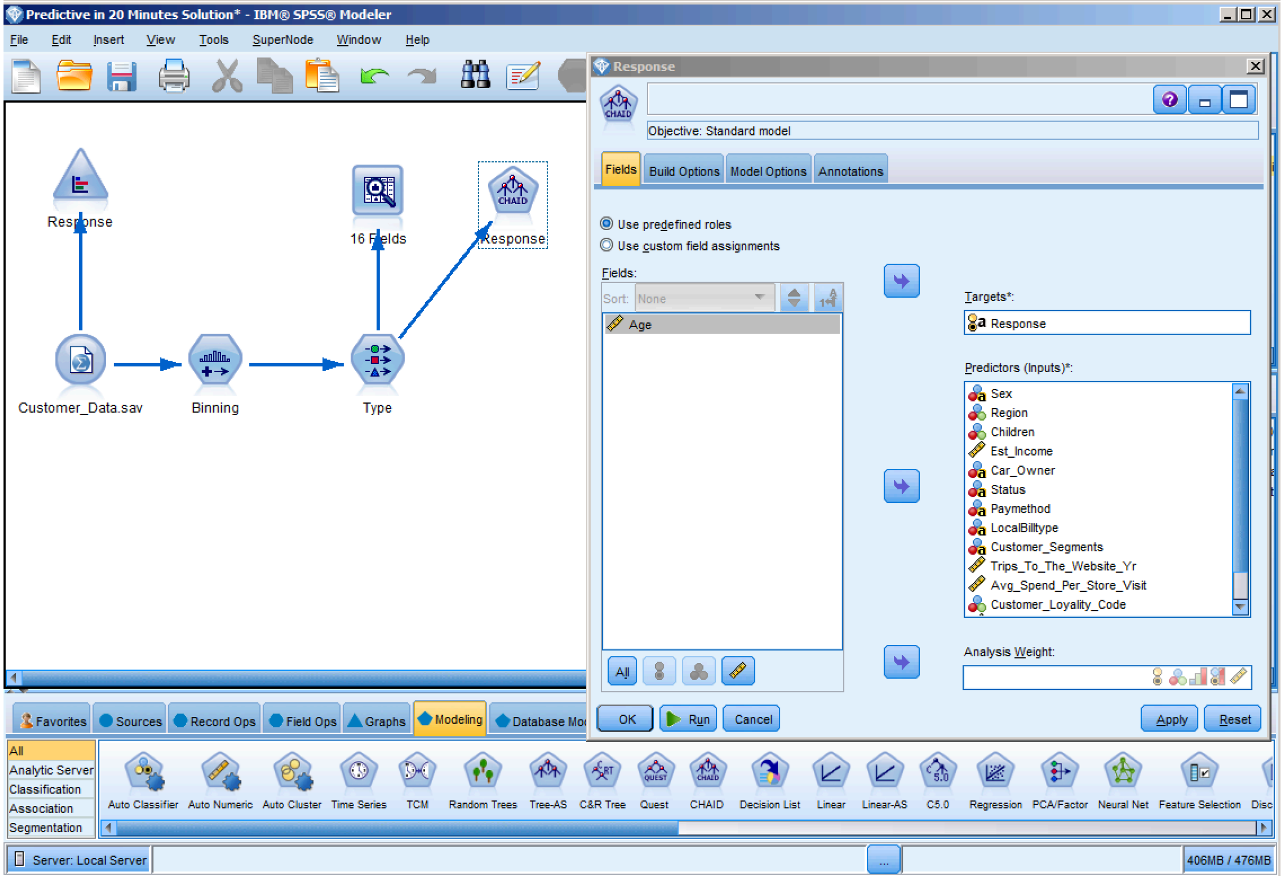


Field	Sample Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique
ID		Continuous	1.000	3825.000	1901.152	1094.709	0.045	--
Sex		Nominal	--	--	--	--	--	2
Region		Nominal	1.000	3.000	--	--	--	3
Children		Nominal	0.000	2.000	--	--	--	3
Est_Income		Continuous	96.330	120000.000	51464.262	30796.681	-0.113	--
Car_Owner		Nominal	--	--	--	--	--	2
Status		Nominal	--	--	--	--	--	3
Paymethod		Nominal	--	--	--	--	--	3

[1] Indicates a multimode result [2] Indicates a sampled result

12. Now that we have explored our data, we can build a model to uncover the key drivers resulting in campaign response. We will do this by connecting a CHAID node from the Modeling palette to the Type node.

Double-click on the CHAID node to review the settings. Since we have already declared Response as our Target, they are predefined. Note that the Build Options tab (not shown) allows a user to select parameters for how data are treated, either automatically or one level at a time. Click Run to execute the model.


Predictive in 20 Minutes Solution* - IBM® SPSS® Modeler

Response

Objective: Standard model

● Use predefined roles
○ Use custom field assignments

Fields:
Sort: None
Age

Targets*:
Response

Predictors (Inputs)*:
Sex
Region
Children
Est_Income
Car_Owner
Status
Paymethod
LocalBilltype
Customer_Segments
Trips_To_The_Website_Yr
Avg_Spend_Per_Store_Visit
Customer_Loyality_Code

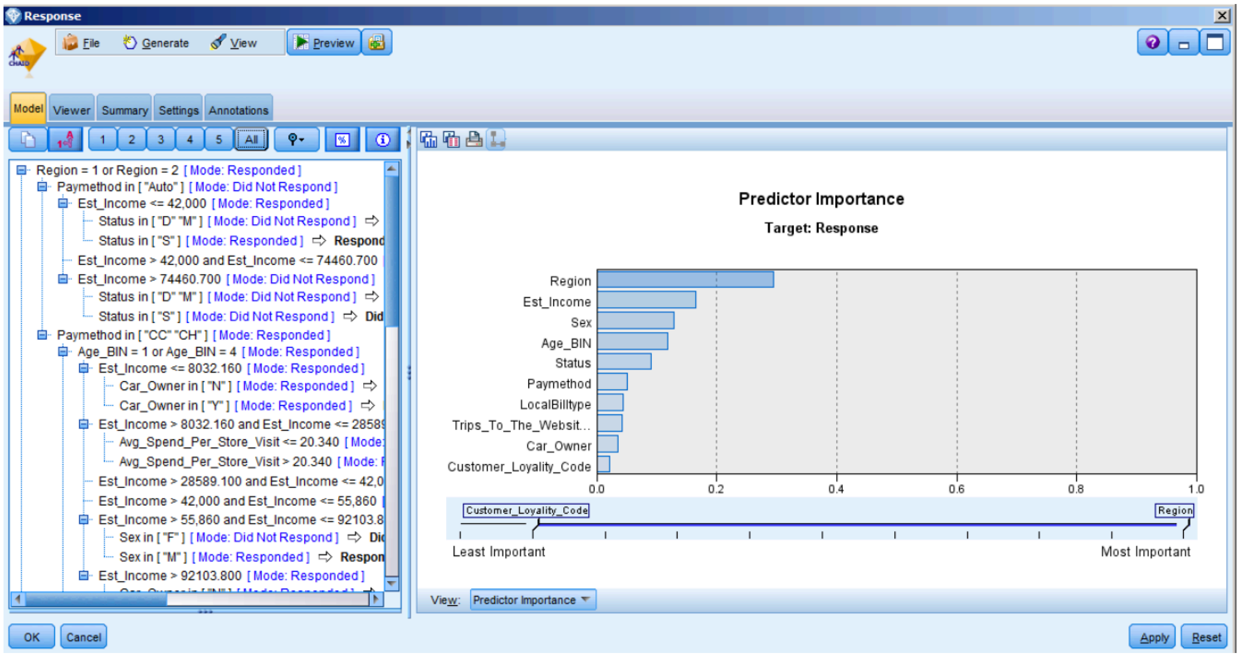
Analysis Weight:

Server: Local Server

406MB / 476MB

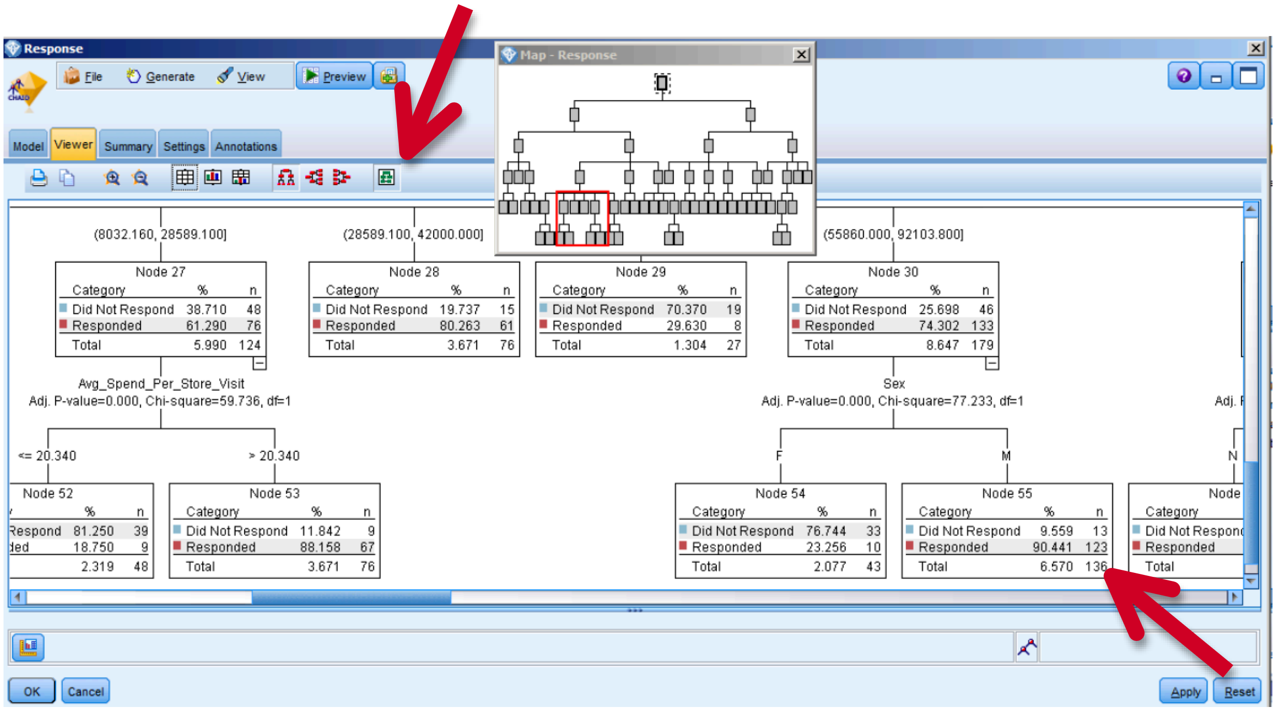
13. The CHAID model is automatically generated and added to our canvas (not shown).

Double-click on the generated model to review the output. In the Model tab of the resulting output, we are provided with a list of the most important predictors to campaign response; the first few being Region, Estimated Income and Sex; as well as rules for the model.



```
Region = 1 or Region = 2 [ Mode: Responded ]
    Paymethod in [ "Auto" ] [ Mode: Did Not Respond ]
        Est_Income <= 42,000 [ Mode: Responded ]
            Status in [ "D" "M" ] [ Mode: Did Not Respond ] ⇨
            Status in [ "S" ] [ Mode: Responded ] ⇨ Respond
        Est_Income > 42,000 and Est_Income <= 74460.700
        Est_Income > 74460.700 [ Mode: Did Not Respond ]
            Status in [ "D" "M" ] [ Mode: Did Not Respond ] ⇨
            Status in [ "S" ] [ Mode: Did Not Respond ] ⇨ Did
    Paymethod in [ "CC" "CH" ] [ Mode: Responded ]
        Age_BIN = 1 or Age_BIN = 4 [ Mode: Responded ]
            Est_Income <= 8032.160 [ Mode: Responded ]
                Car_Owner in [ "N" ] [ Mode: Responded ] ⇨
                Car_Owner in [ "Y" ] [ Mode: Responded ] ⇨
            Est_Income > 8032.160 and Est_Income <= 28589
                Avg_Spend_Per_Store_Visit <= 20.340 [ Mode:
                Avg_Spend_Per_Store_Visit > 20.340 [ Mode: F
            Est_Income > 28589.100 and Est_Income <= 42,0
            Est_Income > 42,000 and Est_Income <= 55,860
            Est_Income > 55,860 and Est_Income <= 92103.8
                Sex in [ "F" ] [ Mode: Did Not Respond ] ⇨ Di
                Sex in [ "M" ] [ Mode: Responded ] ⇨ Respon
            Est_Income > 92103.800 [ Mode: Responded ]
```

14. The Viewer tab displays the resulting tree, reflecting cut points of the important predictors as determined by the model. To navigate the tree map, click on the “show or hide tree map window” icon in the toolbar. Following one branch down, we are able to discover key insights. For example, males who earned an income between $55K - $92K, whose age fell within BINs 1 or 4, who used CC or CH payment methods and lived in Regions 1 or 2; responded to the campaign at very high rates.



(8032.160, 28589.100]

Node 27

Category	%	n
🟦 Did Not Respond	38.710	48
🟥 Responded	61.290	76
Total	5.990	124

Avg_Spend_Per_Store_Visit
Adj. P-value=0.000, Chi-square=59.736, df=1

<= 20.340

Node 52

	%	n
Respond	81.250	39
ded	18.750	9
	2.319	48

> 20.340

Node 53

Category	%	n
🟦 Did Not Respond	11.842	9
🟥 Responded	88.158	67
Total	3.671	76

(28589.100, 42000.000]

Node 28

Category	%	n
🟦 Did Not Respond	19.737	15
🟥 Responded	80.263	61
Total	3.671	76

Node 29

Category	%	n
🟦 Did Not Respond	70.370	19
🟥 Responded	29.630	8
Total	1.304	27

(55860.000, 92103.800]

Node 30

Category	%	n
🟦 Did Not Respond	25.698	46
🟥 Responded	74.302	133
Total	8.647	179

Sex
Adj. P-value=0.000, Chi-square=77.233, df=1

F

Node 54

Category	%	n
🟦 Did Not Respond	76.744	33
🟥 Responded	23.256	10
Total	2.077	43

M

Node 55

Category	%	n
🟦 Did Not Respond	9.559	13
🟥 Responded	90.441	123
Total	6.570	136

Adj. F

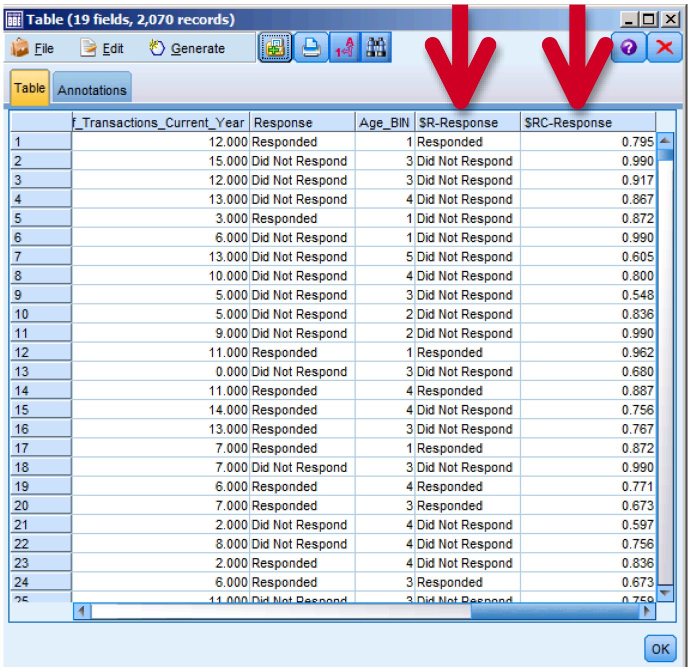
N

Node

Category	%	n
🟦 Did Not Respond		
🟥 Responded		
Total		

15. To view model output, from the Output palette, add a Table node to the canvas and connect it to the CHAID model node.

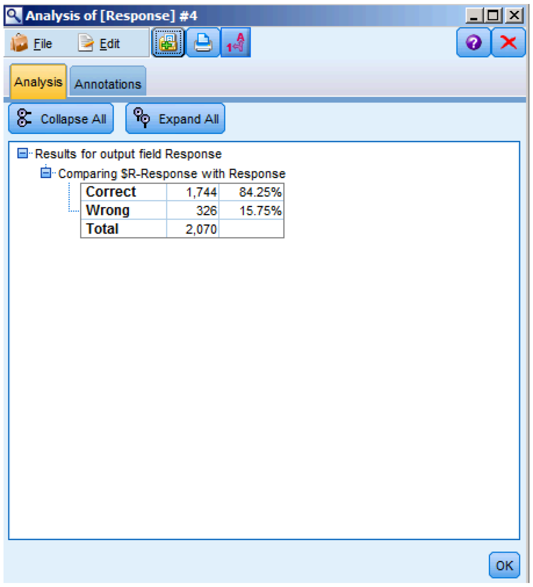
Run the Table by right-clicking and selecting Run, or by using the Run Selection toolbar button. Look at the last two columns of the table. The second to last column contains the predicted response outcomes, which can be compared to the historical outcomes in the fourth to last column; and the last column contains the confidence of that prediction. For example, the first record shows a customer who did, in fact, respond to the campaign. The appended columns show that the model predicted that the customer would respond with 79.5% confidence.



	f_Transactions_Current_Year	Response	Age_BIN	$R-Response	$RC-Response
1	12.000	Responded	1	Responded	0.795
2	15.000	Did Not Respond	3	Did Not Respond	0.990
3	12.000	Did Not Respond	3	Did Not Respond	0.917
4	13.000	Did Not Respond	4	Did Not Respond	0.867
5	3.000	Responded	1	Did Not Respond	0.872
6	6.000	Did Not Respond	1	Did Not Respond	0.990
7	13.000	Did Not Respond	5	Did Not Respond	0.605
8	10.000	Did Not Respond	4	Did Not Respond	0.800
9	5.000	Did Not Respond	3	Did Not Respond	0.548
10	5.000	Did Not Respond	2	Did Not Respond	0.836
11	9.000	Did Not Respond	2	Did Not Respond	0.990
12	11.000	Responded	1	Responded	0.962
13	0.000	Did Not Respond	3	Did Not Respond	0.680
14	11.000	Responded	4	Responded	0.887
15	14.000	Responded	4	Did Not Respond	0.756
16	13.000	Responded	3	Did Not Respond	0.767
17	7.000	Responded	1	Responded	0.872
18	7.000	Did Not Respond	3	Did Not Respond	0.990
19	6.000	Responded	4	Responded	0.771
20	7.000	Responded	3	Responded	0.673
21	2.000	Did Not Respond	4	Did Not Respond	0.597
22	8.000	Did Not Respond	4	Did Not Respond	0.756
23	2.000	Responded	4	Did Not Respond	0.836
24	6.000	Responded	3	Responded	0.673
25	11.000	Did Not Respond	3	Did Not Respond	0.759

16. To see the overall accuracy of the model, select an Analysis node from the Output palette, connect it to the generated model, and select Run.

The resulting output indicates an overall accuracy of 84.25%. That is, the model predicted with 84.25% accuracy which customers responded or did not respond to the campaign.



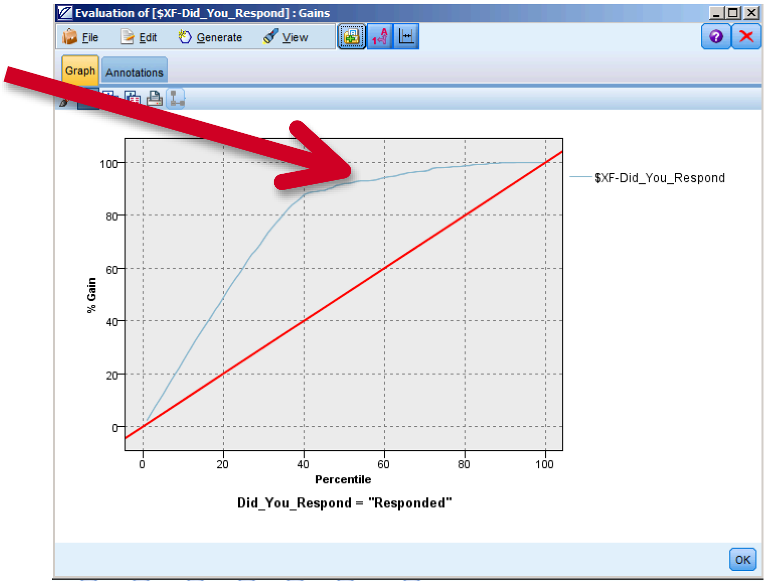
Results for output field Response

Comparing $R-Response with Response

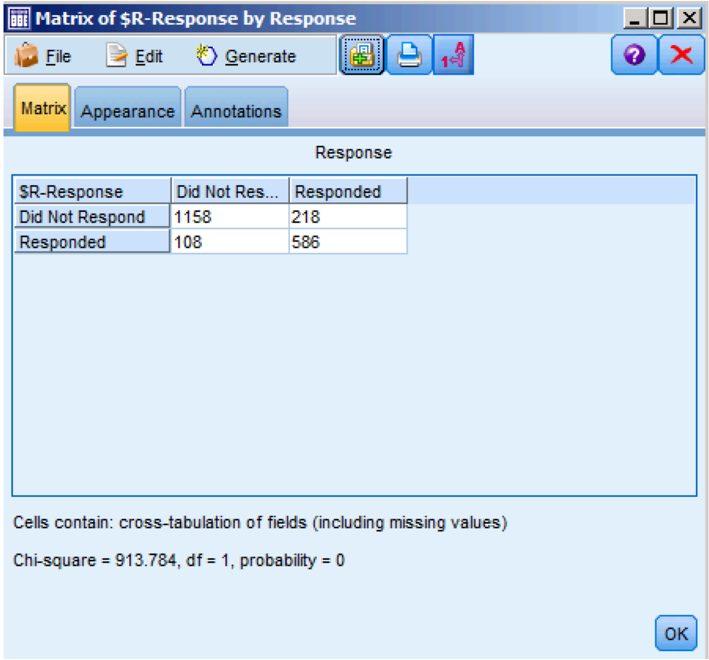
Correct	1,744	84.25%
Wrong	326	15.75%
Total	2,070	

To further evaluate the model, select an Evaluation node from the Graphs palette, connect it to the model node and select Run.

In the resulting gains chart, the red line reflects what you could expect without Predictive Analytics. The blue line; however, reflects the lift in response you could achieve utilizing Predictive Analytics. Therefore, if you were to randomly select 50% of your client base, you could expect to have captured 50% of those likely to respond. By using Predictive Analytics, you can more effectively target those 50% of clients and capture almost 90% of those likely to respond.



17. Finally, to see the relationships between fields, select a Matrix node from the Output palette and connect it to the model node. Using the drop-down menu, choose “$R-Responses” for Rows and “Response” for Columns. Select Run.



Response

$R-Response	Did Not Res...	Responded
Did Not Respond	1158	218
Responded	108	586

Cells contain: cross-tabulation of fields (including missing values)

Chi-square = 913.784, df = 1, probability = 0

Summary

- ✓ Use a data extract from a CRM
- ✓ Prepare data for modeling
- ✓ Define which fields to use
- ✓ Choose the modeling technique
- ✓ Automatically generate a model to identify who has responded
- ✓ Review results

Over the course of the last 20 minutes, we were able to successfully train a model by exploring IBM SPSS Modeler's ability to read in data from a variety of sources, create new fields via data preparation techniques, choose and run a predictive modeling algorithm, and evaluate the results to accurately identify customer response.

Exercise 2: Finding Patterns and Groups

Use Case

Goal: Create segments of customers

Approach:

- Merge disparate data sources, including customer data from a database or CRM
- Define which fields to use
- Automatically generate a model to group customers
- Apply business terms to new customer groups
- Export newly created groups to a database

Why?

- Better customer understanding (demographics, socio-economic, etc.)
- Tailored messages for each group/segment
- Personal and more relevant for consumers

Customer Reference

A US cable television network turns on the insights with an analytics solution that predicts success of new shows six weeks in advance.

Business challenge: This cable television network faces the challenge of managing huge volumes of information. Previously the network's research team spent a significant amount of time processing data on spreadsheets rather than analyzing it, and based decisions on a combination of experience and instinct. The company needed a large-scale analytics solution to organize this wealth of data, make sense of it, and provide answers and actionable insights.

The transformation: The solution combines television ratings data with information gathered minute by minute and viewer by viewer from a variety of channels and other sources to determine who's watching and why. Then it centralizes the data and makes it available for in-depth, predictive analytics. With insights into audience preferences gained from sophisticated statistical models, including intelligent segmentation, the network can optimize advertising revenue and viewership like never before.

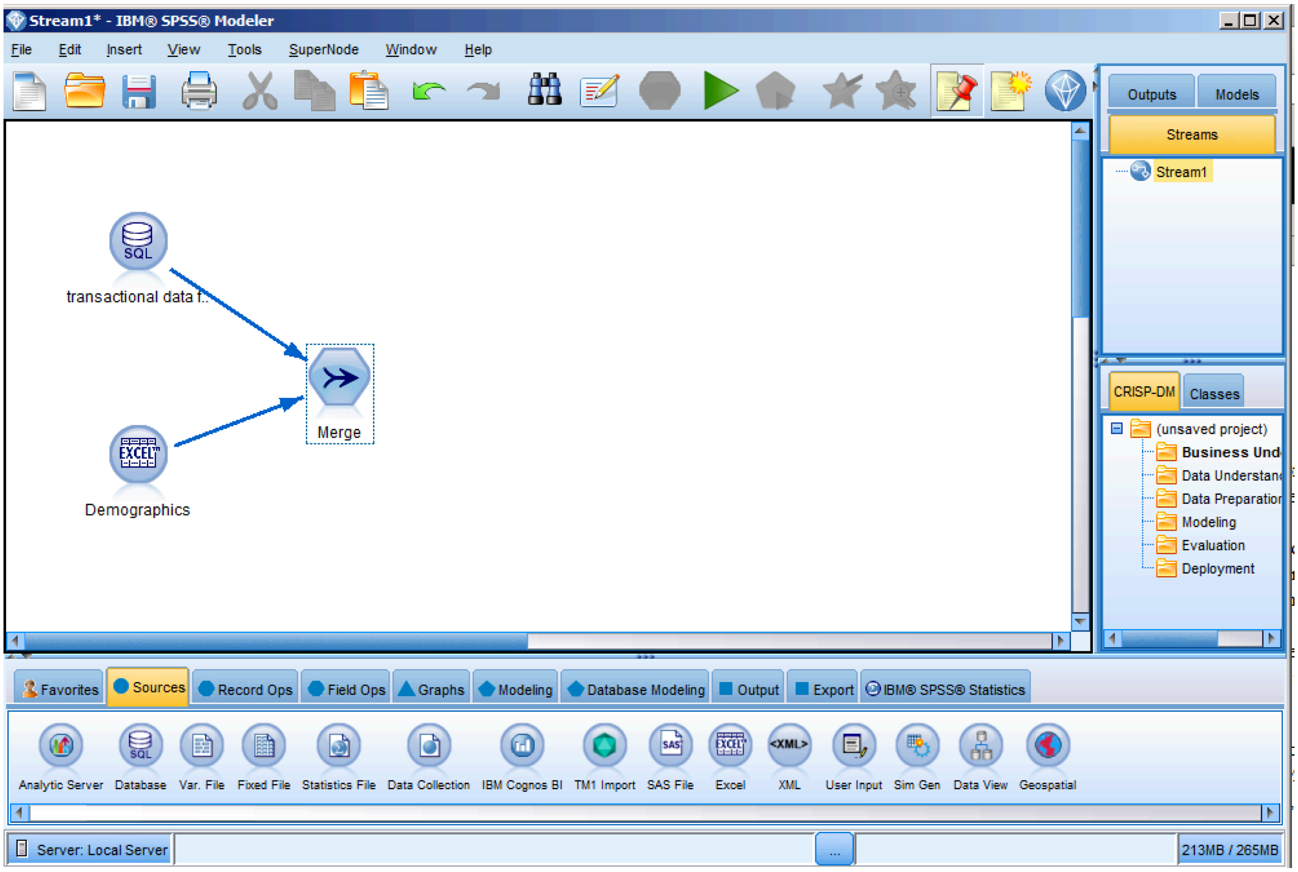
IBM's implemented solution:
Accelerates analytics by extracting insights from billions of rows of audience data in seconds, instead of days.
Triples views of video-on-demand service through data-driven marketing.
Predicts success of a new show six weeks in advance of its release and adjusts marketing accordingly.

"A single day of analysis work enabled the network to design a campaign that increased the consumption of its video on demand service. Previously, that analysis would have taken weeks."

Finding Patterns and Groups

1. Open the Customer Segmentation.str file from the workshop directory. In IBM SPSS Modeler, click on File, Open stream, and then navigate to: C:\Modeler Workshop\Segmentation\Customer Segmentation.str.

   Either double-click on Customer Segmentation.str, or select it and then click on Open.



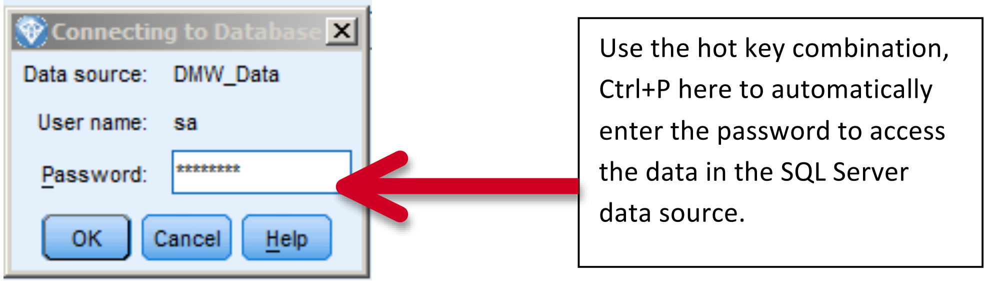
You might recall that in the Predictive in 20 Minutes exercise there was a field assigning each customer to a marketing cluster or segment. We're going to take a step back in our story to take a deeper look at how that is accomplished in IBM SPSS Modeler.

We will be joining two data files together; one file contains customer transaction data, and exists in a database, while the other file is an Excel file and contains customer demographics. Though the files are in different formats, the Merge node in IBM SPSS Modeler can very easily join the files without first requiring the analyst to translate one file and/or the other into the same format.

To get us started quickly, the two source nodes on the canvas, and the merge node used to join them together, have already been configured and connected for you.

2. From the Output palette, add a Table node to the canvas and connect it to the Merge node.

3. Once connected, right click on the Table node and select Run to review the data after having been merged. If prompted for a password, use “Ctrl+P” to automatically populate the password.



Connecting to Database

Data source: DMW_Data

User name: sa

Password: ********

Use the hot key combination, Ctrl+P here to automatically enter the password to access the data in the SQL Server data source.

The previous requirement to provide a password only needs to be done once during this workshop; after logging into the SQL Server data source now, you will have continued access to that data source throughout the workshop and will not be prompted for that again.

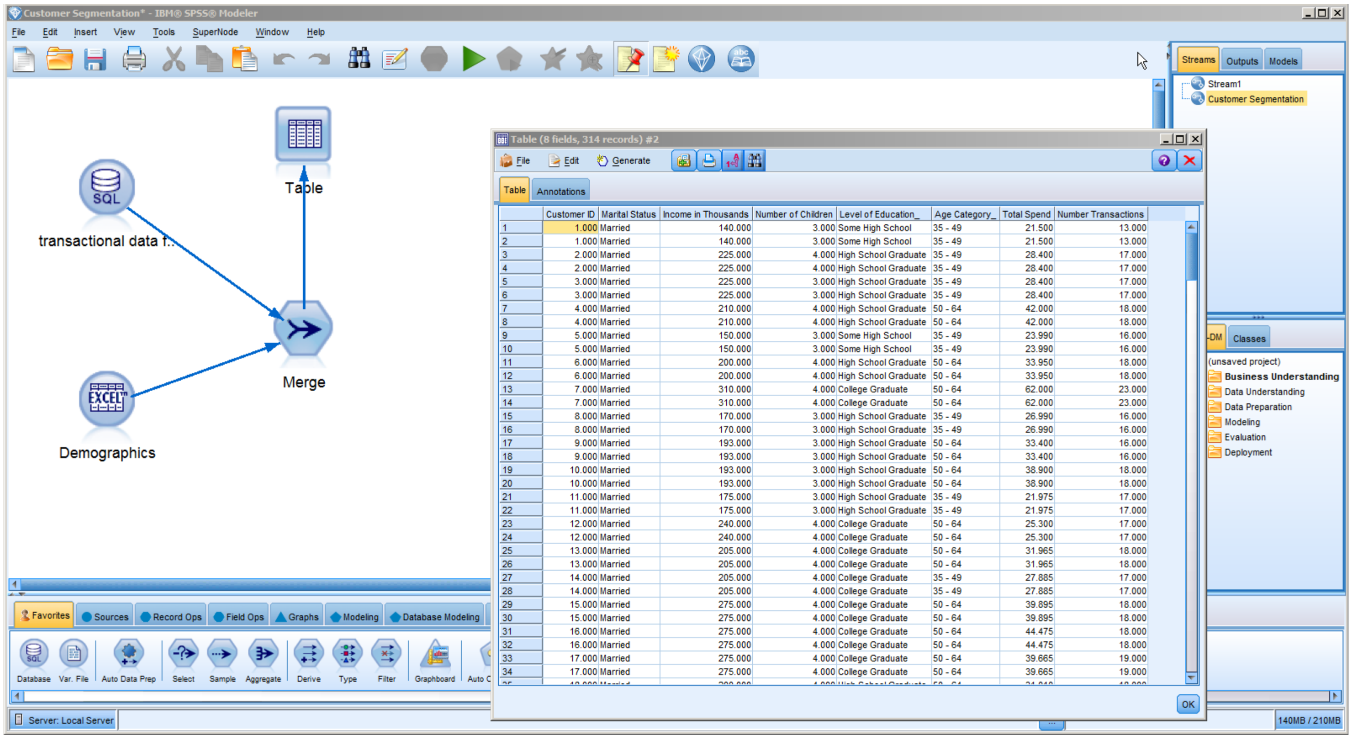
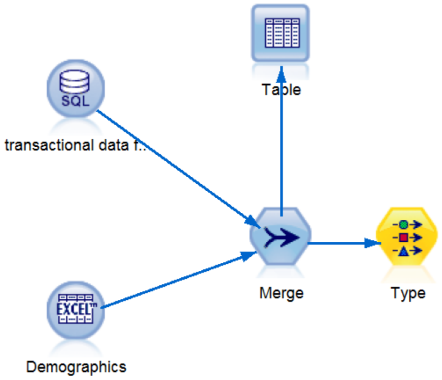
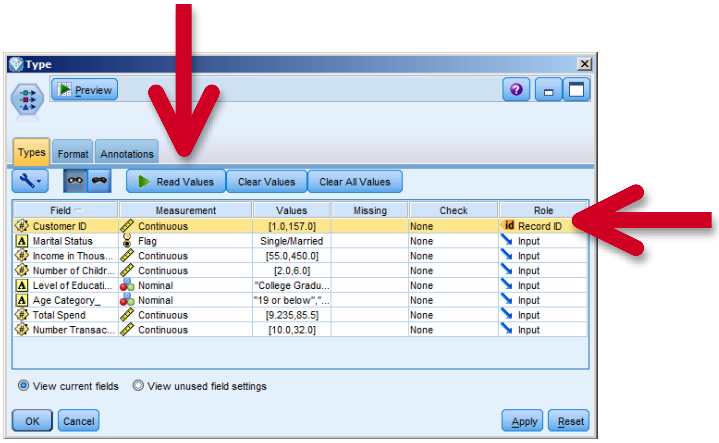


Table (8 fields, 314 records) #2

	Customer ID	Marital Status	Income in Thousands	Number of Children	Level of Education_	Age Category_	Total Spend	Number Transactions
1	1.000	Married	140.000	3.000	Some High School	35 - 49	21.500	13.000
2	1.000	Married	140.000	3.000	Some High School	35 - 49	21.500	13.000
3	2.000	Married	225.000	4.000	High School Graduate	35 - 49	28.400	17.000
4	2.000	Married	225.000	4.000	High School Graduate	35 - 49	28.400	17.000
5	3.000	Married	225.000	3.000	High School Graduate	35 - 49	28.400	17.000
6	3.000	Married	225.000	3.000	High School Graduate	35 - 49	28.400	17.000
7	4.000	Married	210.000	4.000	High School Graduate	50 - 64	42.000	18.000
8	4.000	Married	210.000	4.000	High School Graduate	50 - 64	42.000	18.000
9	5.000	Married	150.000	3.000	Some High School	35 - 49	23.990	16.000
10	5.000	Married	150.000	3.000	Some High School	35 - 49	23.990	16.000
11	6.000	Married	200.000	4.000	High School Graduate	50 - 64	33.950	18.000
12	6.000	Married	200.000	4.000	High School Graduate	50 - 64	33.950	18.000
13	7.000	Married	310.000	4.000	College Graduate	50 - 64	62.000	23.000
14	7.000	Married	310.000	4.000	College Graduate	50 - 64	62.000	23.000
15	8.000	Married	170.000	3.000	High School Graduate	35 - 49	26.990	16.000
16	8.000	Married	170.000	3.000	High School Graduate	35 - 49	26.990	16.000
17	9.000	Married	193.000	3.000	High School Graduate	50 - 64	33.400	16.000
18	9.000	Married	193.000	3.000	High School Graduate	50 - 64	33.400	16.000
19	10.000	Married	193.000	3.000	High School Graduate	50 - 64	38.900	18.000
20	10.000	Married	193.000	3.000	High School Graduate	50 - 64	38.900	18.000
21	11.000	Married	175.000	3.000	High School Graduate	35 - 49	21.975	17.000
22	11.000	Married	175.000	3.000	High School Graduate	35 - 49	21.975	17.000
23	12.000	Married	240.000	4.000	College Graduate	50 - 64	25.300	17.000
24	12.000	Married	240.000	4.000	College Graduate	50 - 64	25.300	17.000
25	13.000	Married	205.000	4.000	College Graduate	50 - 64	31.965	18.000
26	13.000	Married	205.000	4.000	College Graduate	50 - 64	31.965	18.000
27	14.000	Married	205.000	4.000	College Graduate	35 - 49	27.885	17.000
28	14.000	Married	205.000	4.000	College Graduate	35 - 49	27.885	17.000
29	15.000	Married	275.000	4.000	College Graduate	50 - 64	39.895	18.000
30	15.000	Married	275.000	4.000	College Graduate	50 - 64	39.895	18.000
31	16.000	Married	275.000	4.000	College Graduate	50 - 64	44.475	18.000
32	16.000	Married	275.000	4.000	College Graduate	50 - 64	44.475	18.000
33	17.000	Married	275.000	4.000	College Graduate	50 - 64	39.665	19.000
34	17.000	Married	275.000	4.000	College Graduate	50 - 64	39.665	19.000

4. From the Field Ops palette at the bottom of the screen, select a Type node and attach it to the Merge node. Double-click on the Type node to edit it. Click on the Read Values push button, and change the role of “Customer ID” to “Record ID” (shown below). Once done, click on the OK button.



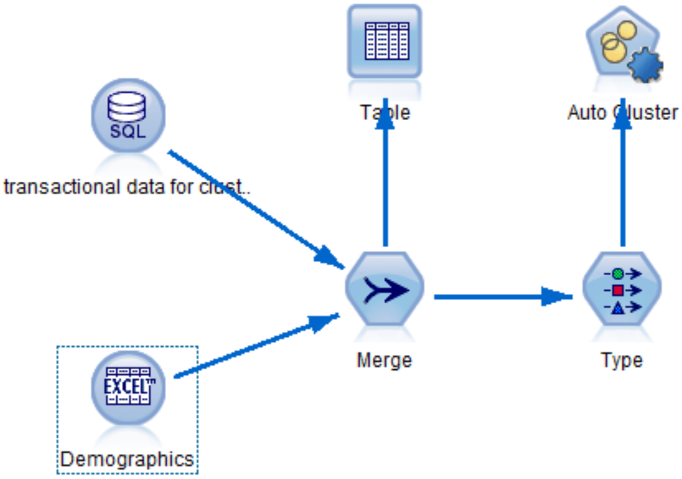


Field	Measurement	Values	Missing	Check	Role
Customer ID	Continuous	[1.0,157.0]		None	Record ID
Marital Status	Flag	Single/Married		None	Input
Income in Thous...	Continuous	[55.0,450.0]		None	Input
Number of Childr...	Continuous	[2.0,6.0]		None	Input
Level of Educati...	Nominal	"College Gradu...		None	Input
Age Category_	Nominal	"19 or below","...		None	Input
Total Spend	Continuous	[9.235,85.5]		None	Input
Number Transac...	Continuous	[10.0,32.0]		None	Input

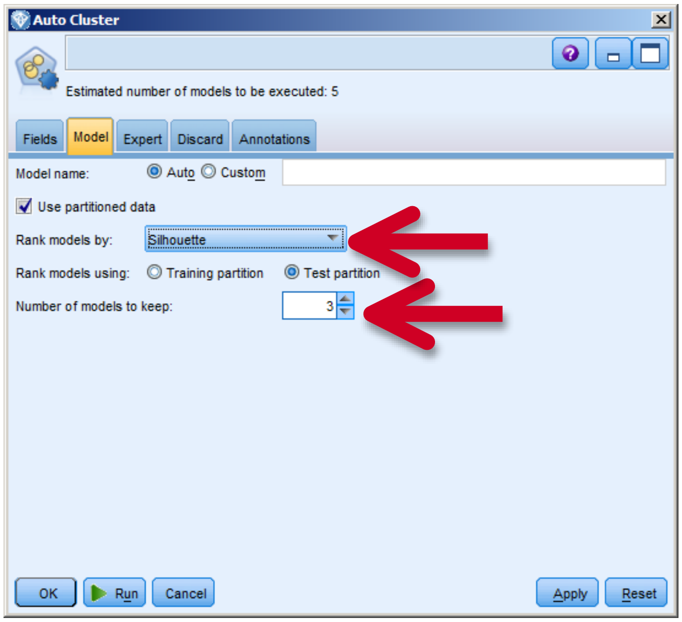
● View current fields ○ View unused field settings

At this point we are ready to cluster the cases into segments. For this we will use the ‘Auto Cluster’ node. The Auto Cluster node allows you to try all of the clustering algorithms and, at your discretion, any or all of their parameters. It builds all of the models you specify and shows you the best models (3 is the default) to use with your data. So, in one step, you will have the best model(s) without having to know or guess which might work for you. This is also a nice way to see how other modeling algorithms will perform. The same holds for the other auto modeling nodes that address classification and numeric modeling techniques.

5. From the Modeling palette select the Auto Cluster modeling node and attach it to the Type node.



Double-click on the Auto Cluster node to edit it.



Estimated number of models to be executed: 5

Model name: ● Auto ○ Custom

☑ Use partitioned data

Rank models by: Silhouette

Rank models using: ○ Training partition ● Test partition

Number of models to keep: 3

On the 'Model' tab, there are various ways of ranking the quality of the models that are built.
Keep the default method, 'Silhouette'.

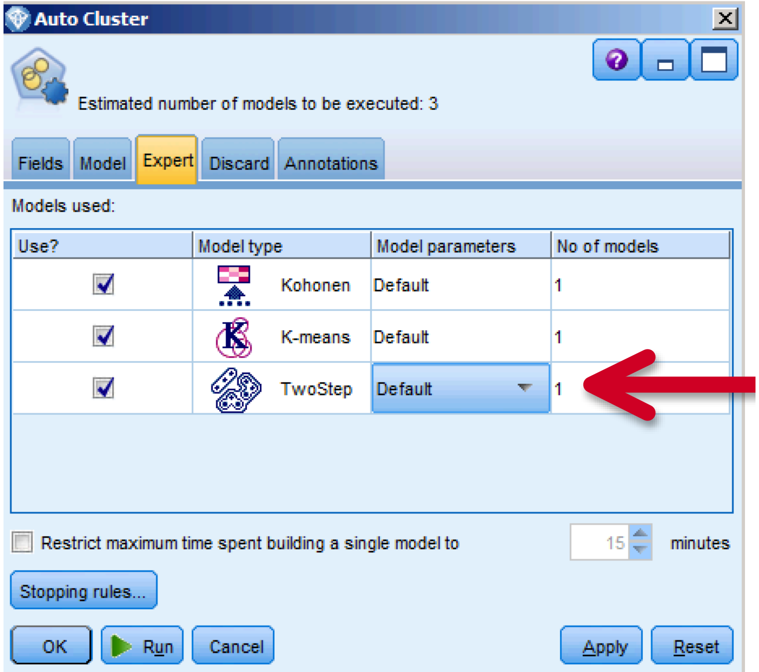
Also keep the default number of models to keep, 3. This means that the 3 best models, based on their silhouette measure, will be retained for our use.

Click on the 'Expert' tab.

You notice that there are 3 clustering algorithms in the list. Model parameters for each can be specified.

For example, click on the word **Default** to the right of TwoStep. Then click on **Specify** (not shown).

The dialog box for specifying the parameters for the TwoStep clustering algorithm with appear as shown in the next section.



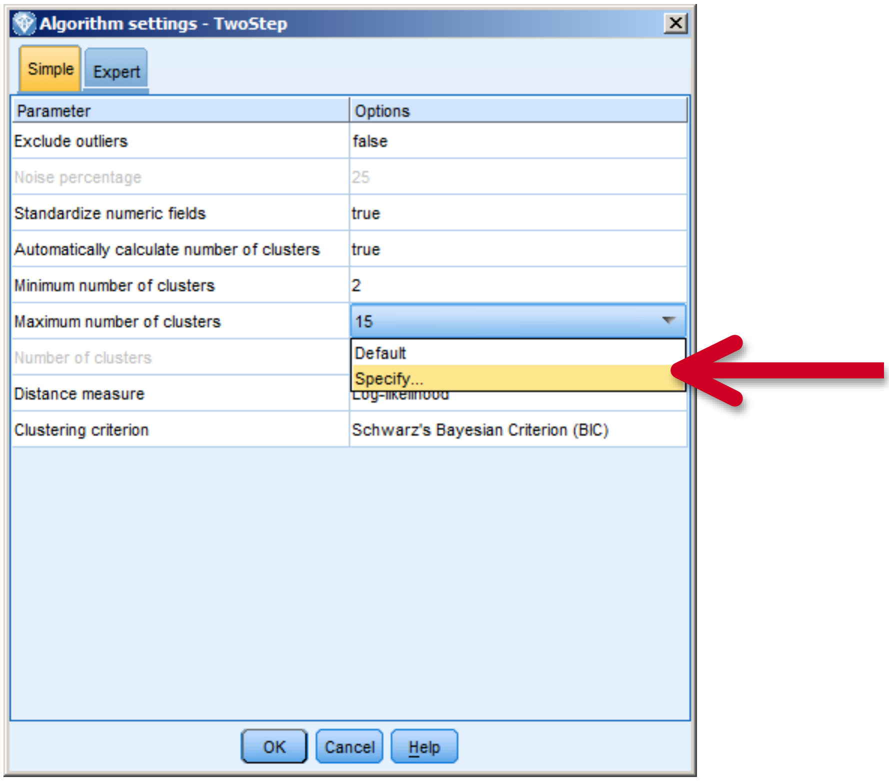
Estimated number of models to be executed: 3

Models used:

Use?	Model type	Model parameters	No of models
☑	Kohonen	Default	1
☑	K-means	Default	1
☑	TwoStep	Default	1

☐ Restrict maximum time spent building a single model to 15 minutes

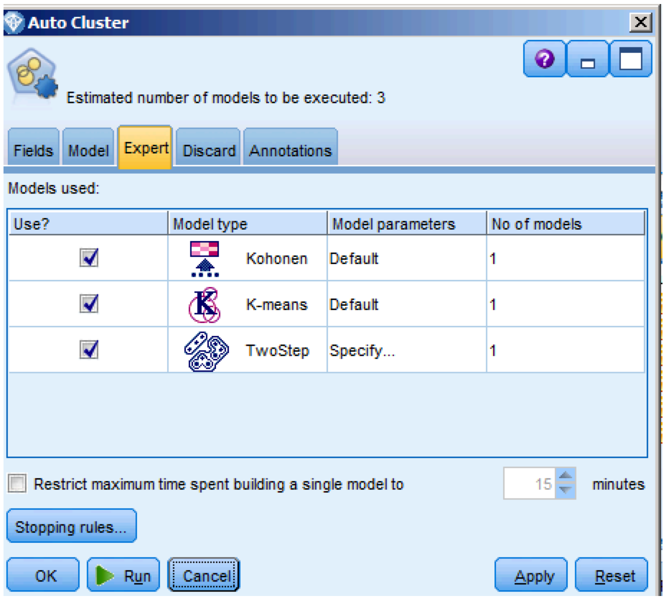
The ‘Maximum number of clusters’ is 15. Click on the number **15**, and click **Specify.**



Parameter	Options
Exclude outliers	false
Noise percentage	25
Standardize numeric fields	true
Automatically calculate number of clusters	true
Minimum number of clusters	2
Maximum number of clusters	15
Number of clusters	Default Specify...
Distance measure	Log-likelihood
Clustering criterion	Schwarz's Bayesian Criterion (BIC)

In the Parameter editor, add 4 to the list. Then delete 15 by clicking on it and then clicking the red X (not shown).

Click OK in the parameter editor and click OK in the Algorithm settings.



Estimated number of models to be executed: 3

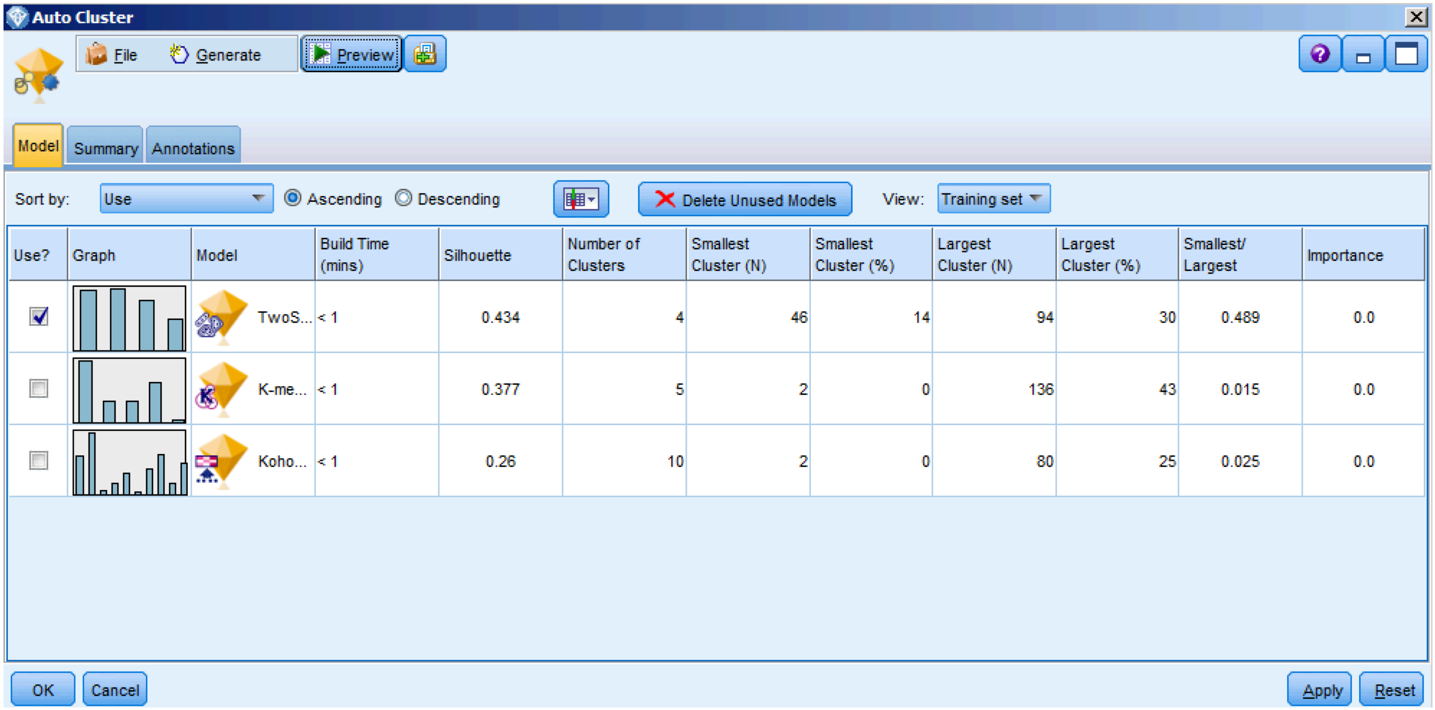
Models used:

Use?	Model type	Model parameters	No of models
☑	Kohonen	Default	1
☑	K-means	Default	1
☑	TwoStep	Specify...	1

□ Restrict maximum time spent building a single model to 15 minutes

Click Run in the Auto Cluster dialog.

6. Double-click on the model nugget (not shown) to view the results of the auto-clustering analysis.



Sort by: Use ● Ascending ○ Descending

View: Training set

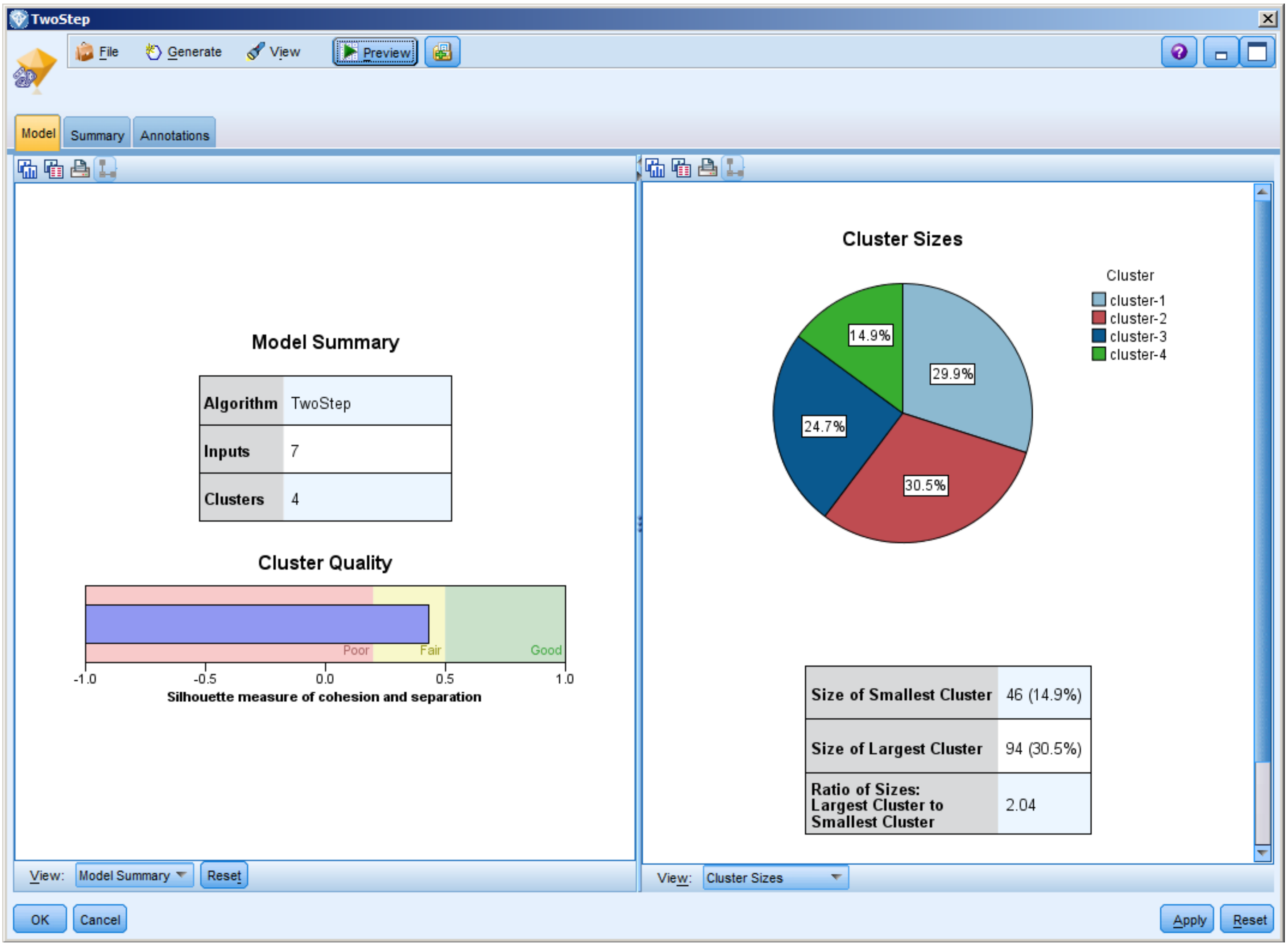
Use?	Graph	Model	Build Time (mins)	Silhouette	Number of Clusters	Smallest Cluster (N)	Smallest Cluster (%)	Largest Cluster (N)	Largest Cluster (%)	Smallest/ Largest	Importance
☑		TwoS...	< 1	0.434	4	46	14	94	30	0.489	0.0
□		K-me...	< 1	0.377	5	2	0	136	43	0.015	0.0
□		Koho...	< 1	0.26	10	2	0	80	25	0.025	0.0

This is where you see the 3 'best' models for segmenting the data. The list is in descending order by Silhouette measure like we specified. There are other important statistics about each model in the table.

The results show that the TwoStep algorithm has the best Silhouette measure followed by the K-means and Kohonen models, which were discarded.

The check boxes to the left indicate that the TwoStep model will be used since it was ranked highest according to our ranking criterion. With cluster models, only one can be selected at a time; but you could choose to use any of the others by clicking the check box. For our exercise we will stay with the TwoStep model.

7. Double-click the first model nugget labeled 'TwoStep 1' to see the results of the Two Step cluster analysis.



Model Summary

Algorithm	TwoStep
Inputs	7
Clusters	4

Cluster Quality

Cluster Sizes

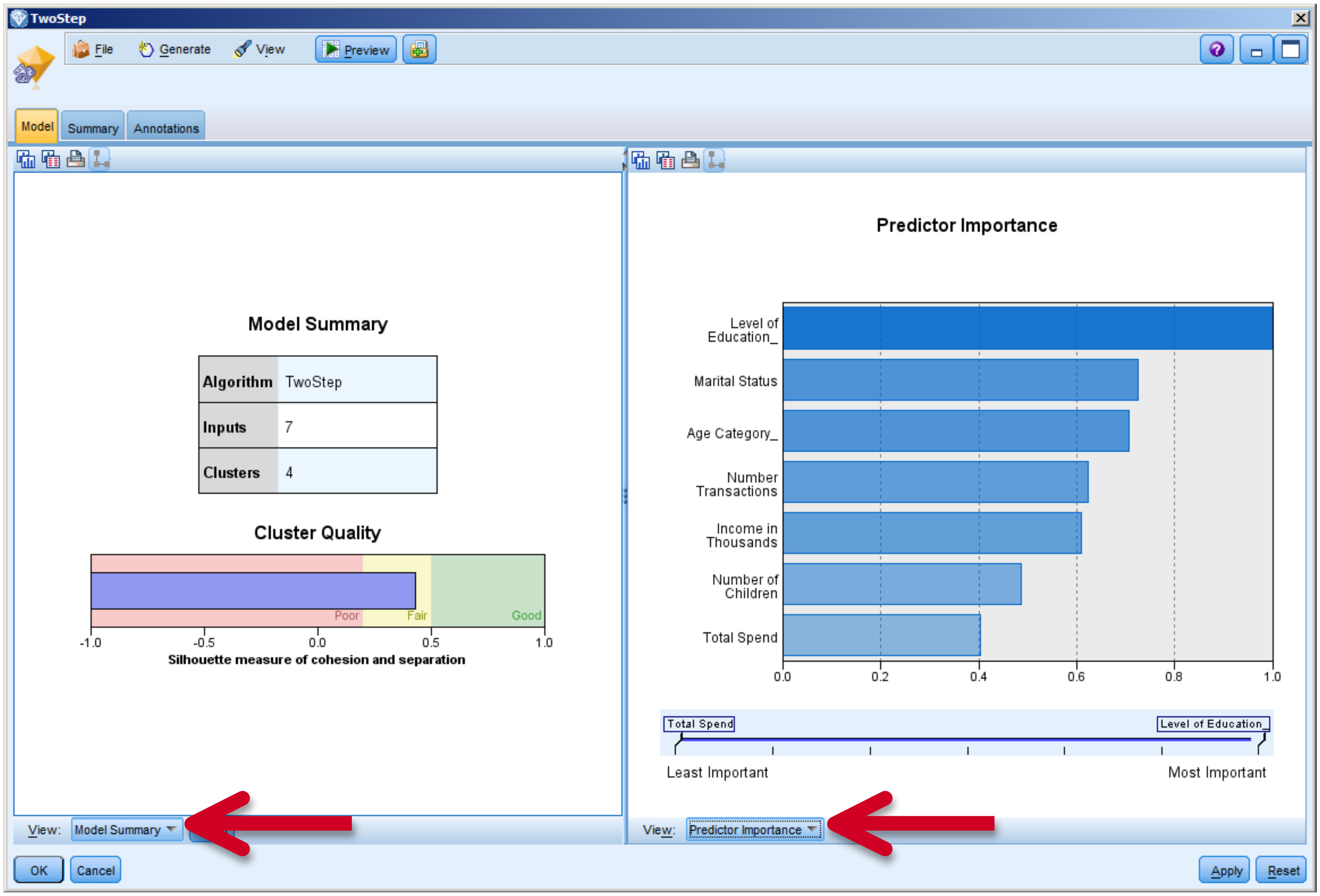
Size of Smallest Cluster	46 (14.9%)
Size of Largest Cluster	94 (30.5%)
Ratio of Sizes: Largest Cluster to Smallest Cluster	2.04

We specified a range of 2 to 4 clusters; and the Two Step clustering engine resolved into 4 clusters.

Looking at the Cluster Quality measure in the left panel, we see that the Silhouette measure (which is a measure of the clusters' internal cohesion AND how well they exclude dissimilar cases) is fair, with a value of just under 0.5. Such results are common, but may also suggest that fewer and/or other variables might be needed to increase the Silhouette value.

On the right side of the viewer is a pie chart illustrating the cluster sizes.

8. From the drop-down menu in the right viewer select Predictor Importance.



Model Summary

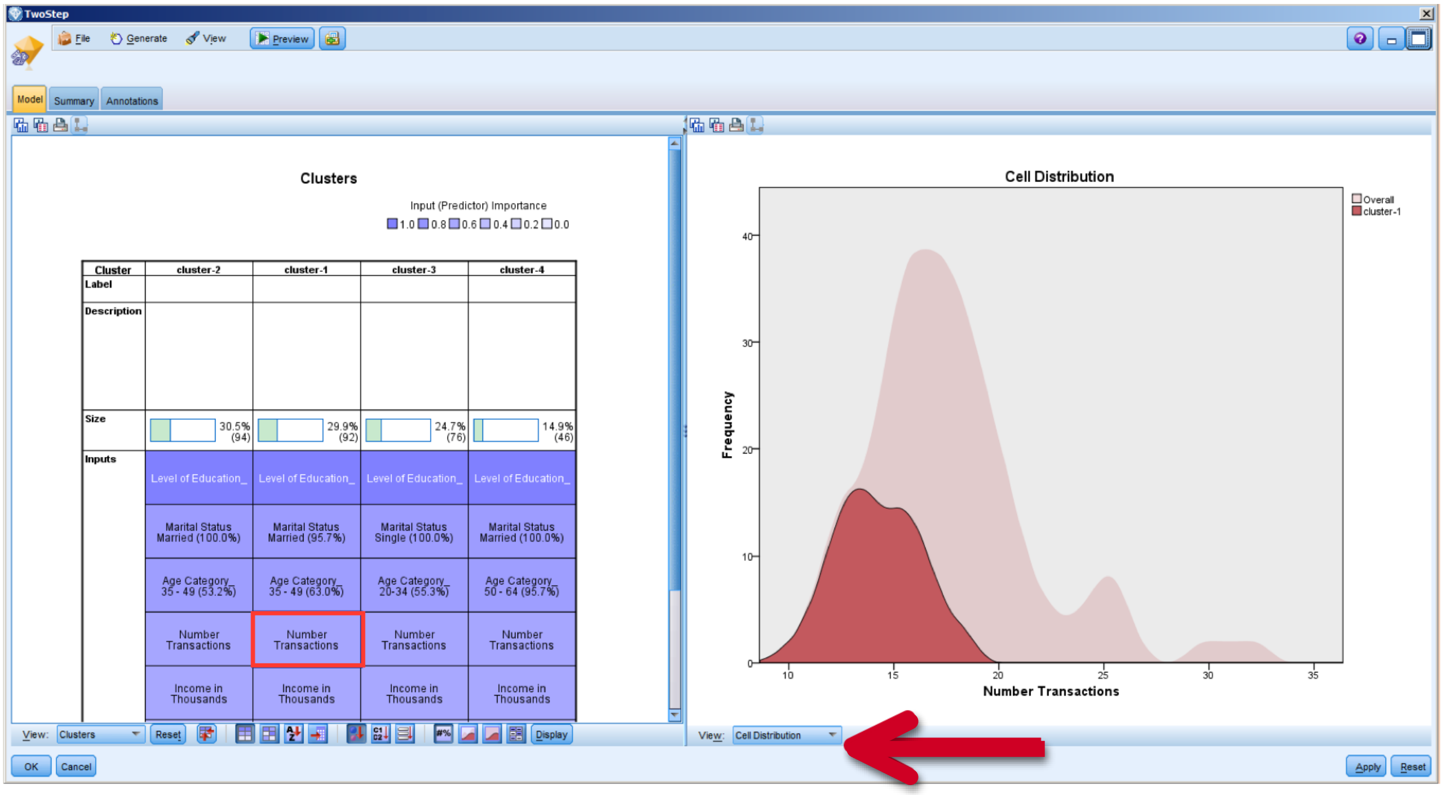
Algorithm	TwoStep
Inputs	7
Clusters	4

View: Model Summary

View: Predictor Importance

Now the right side of the viewer displays a graph with the variables ranked in order of importance for cluster definition. We can see that Level of Education is the most important variable, followed by Marital Status.

9. Now, from the drop-down menu in the left viewer, select Clusters; from the drop-down menu in the right viewer, select Cell Distribution (shown below in red).



View: Clusters

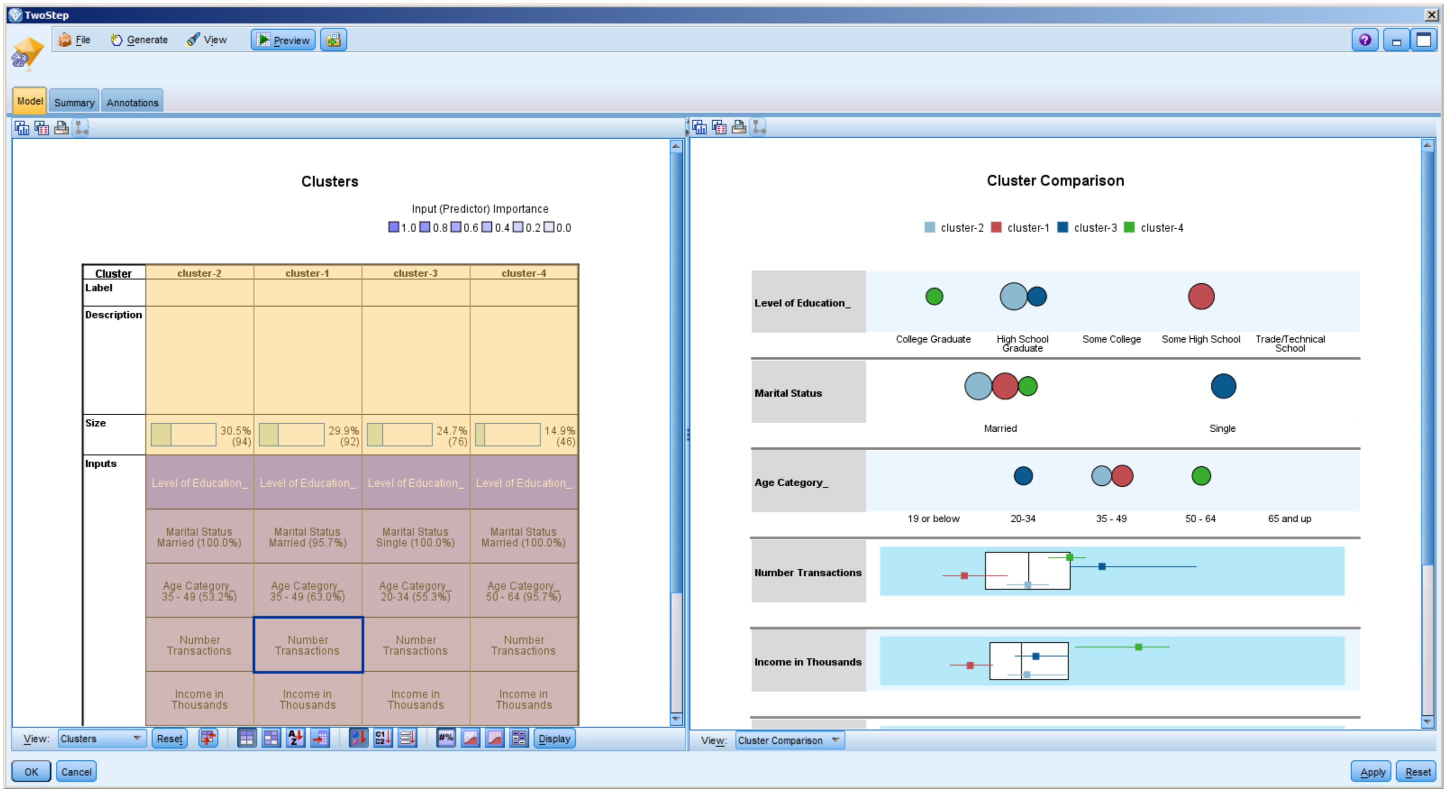
Cluster	cluster-2	cluster-1	cluster-3	cluster-4
Label				
Description				
Size	30.5% (94)	29.9% (92)	24.7% (76)	14.9% (46)
Inputs	Level of Education_	Level of Education_	Level of Education_	Level of Education_
	Marital Status Married (100.0%)	Marital Status Married (95.7%)	Marital Status Single (100.0%)	Marital Status Married (100.0%)
	Age Category_ 35 - 49 (53.2%)	Age Category_ 35 - 49 (63.0%)	Age Category_ 20-34 (55.3%)	Age Category_ 50 - 64 (95.7%)
	Number Transactions	Number Transactions	Number Transactions	Number Transactions
	Income in Thousands	Income in Thousands	Income in Thousands	Income in Thousands

View: Cell Distribution

The left panel of the Viewer displays the clusters in order of their size, left to right. The darkness of the shading of each variable indicates its importance in cluster definition; the lighter the shading, the less important is the variable in defining the clusters.

10. Click on any cell in the grid in the left panel to view, in the right panel, how a cluster distribution compares to the remaining clusters.

11. Click in the left panel on the heading for Cluster-2 and then hold the shift key on your keyboard and simultaneously click on the heading for cluster-4. This selects the entire table. The right panel will display the Cluster Comparison view, which displays the variable distributions relative to each selected cluster.



Clusters

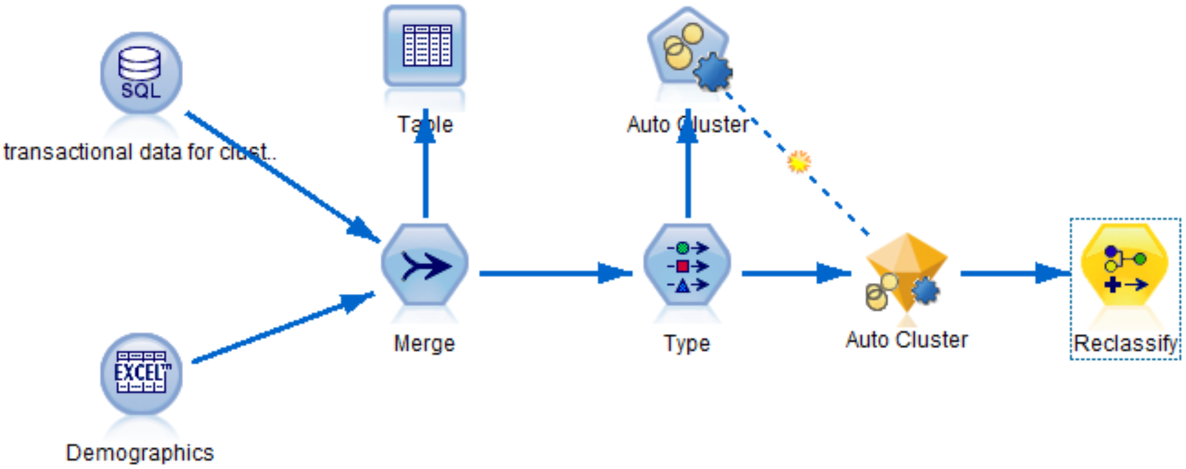
Input (Predictor) Importance
□1.0 □0.8 □0.6 □0.4 □0.2 □0.0

Cluster	cluster-2	cluster-1	cluster-3	cluster-4
Label				
Description				
Size	30.5% (94)	29.9% (92)	24.7% (76)	14.9% (46)
Inputs	Level of Education_	Level of Education_	Level of Education_	Level of Education_
	Marital Status Married (100.0%)	Marital Status Married (95.7%)	Marital Status Single (100.0%)	Marital Status Married (100.0%)
	Age Category_ 35 - 49 (53.2%)	Age Category_ 35 - 49 (63.0%)	Age Category_ 20-34 (55.3%)	Age Category_ 50 - 64 (95.7%)
	Number Transactions	Number Transactions	Number Transactions	Number Transactions
	Income in Thousands	Income in Thousands	Income in Thousands	Income in Thousands

Cluster Comparison

12. After reviewing the Cluster Viewer, click on the red X to close and return to the Modeler workbench.

13. From the Field Ops palette, drag a Reclassify node onto the canvas and connect it to the model node (nugget).

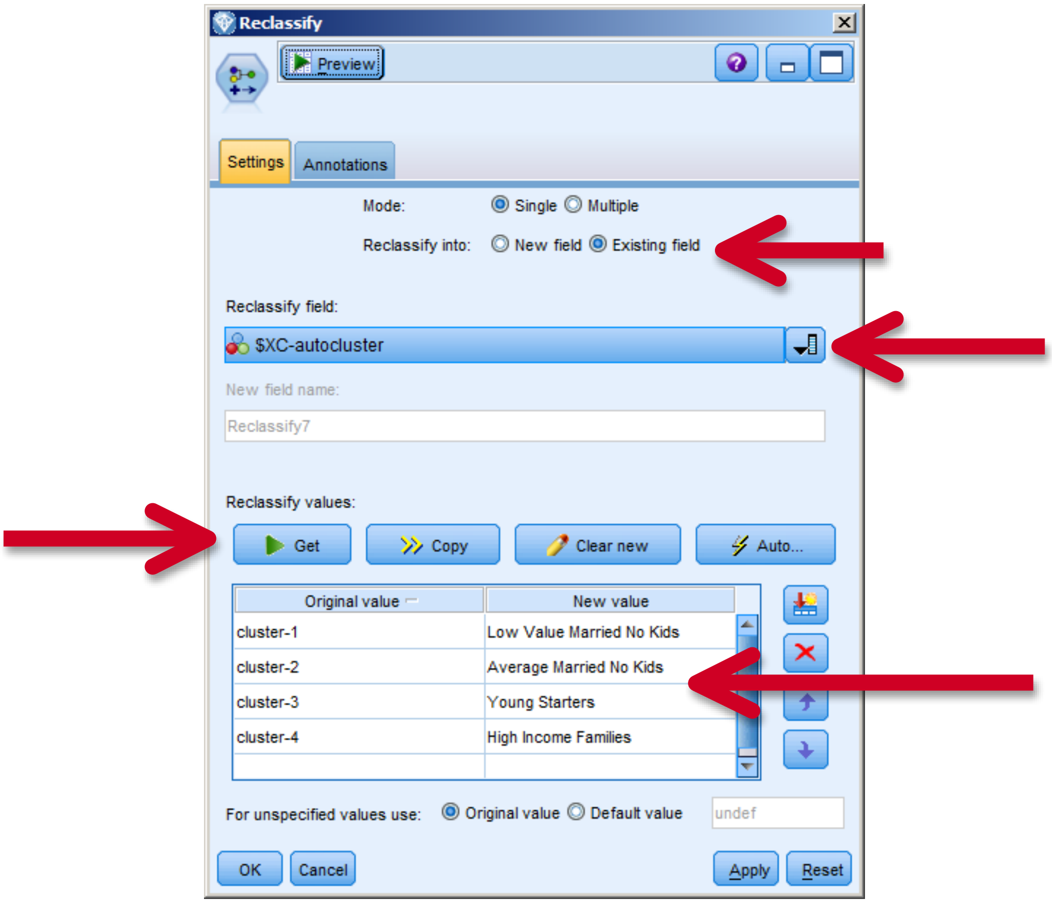


14. Double-click on the Reclassify node to edit it.

In the Settings tab of the Reclassify dialog, use the drop-down menu to select the variable **$XC-autocluster** as the field to be reclassified.

Click “Existing Field” just above so that it doesn’t make a new field.

Click the “**Get**” button to populate the ‘Original value’ column for you. Enter the new values on the right, which better describe the clusters. An example is shown below, but you can enter your own labels as desired. Once completed, click OK.



Mode: ● Single ○ Multiple

Reclassify into: ○ New field ● Existing field

Reclassify field: $XC-autocluster

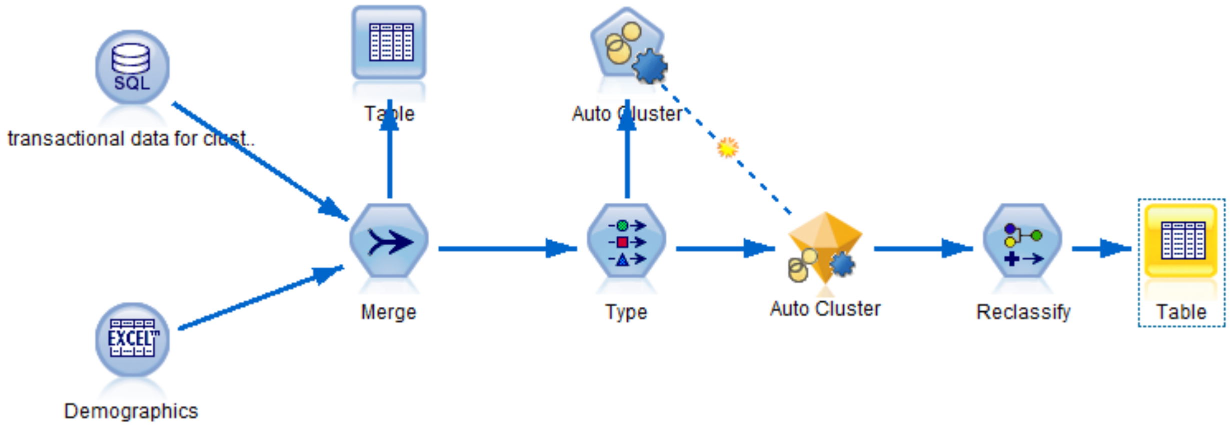
New field name: Reclassify7

Reclassify values:

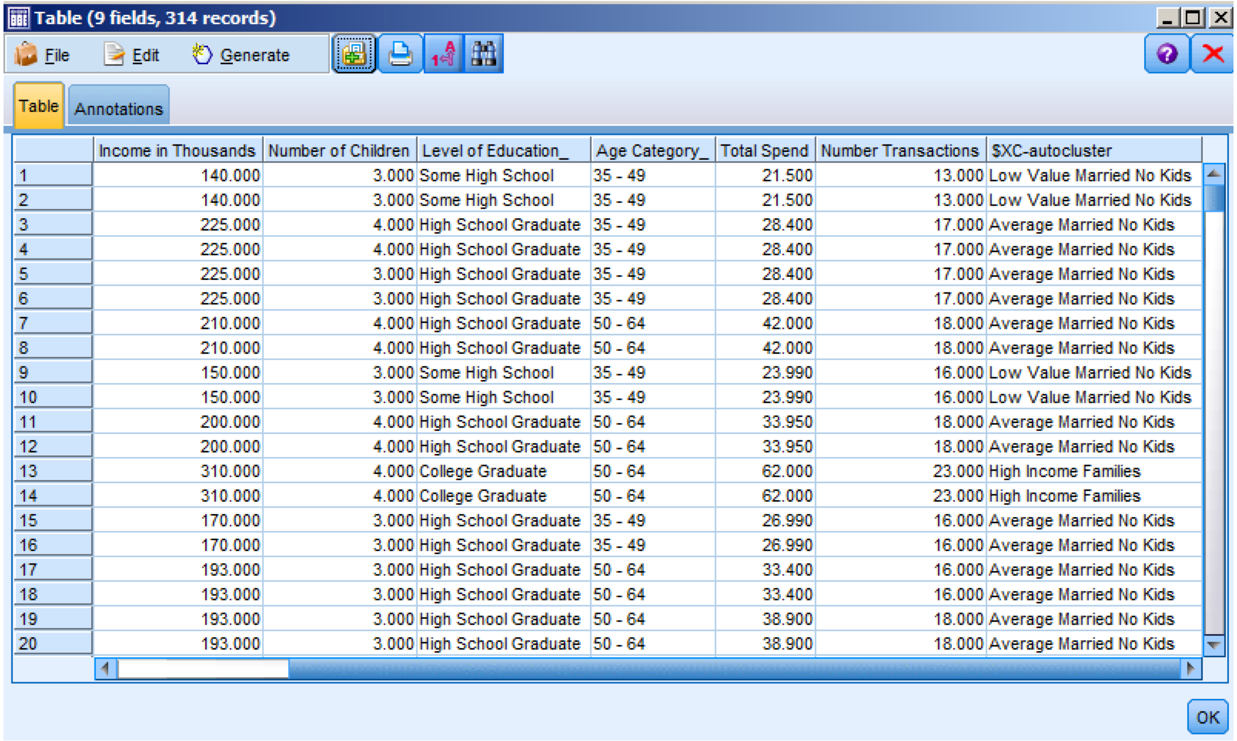
Original value	New value
cluster-1	Low Value Married No Kids
cluster-2	Average Married No Kids
cluster-3	Young Starters
cluster-4	High Income Families

For unspecified values use: ● Original value ○ Default value undef

15. From the Output palette, drag a Table node onto the workbench and connect it to the Reclassify node. Once connected, right-click on the Table node and Select Run.

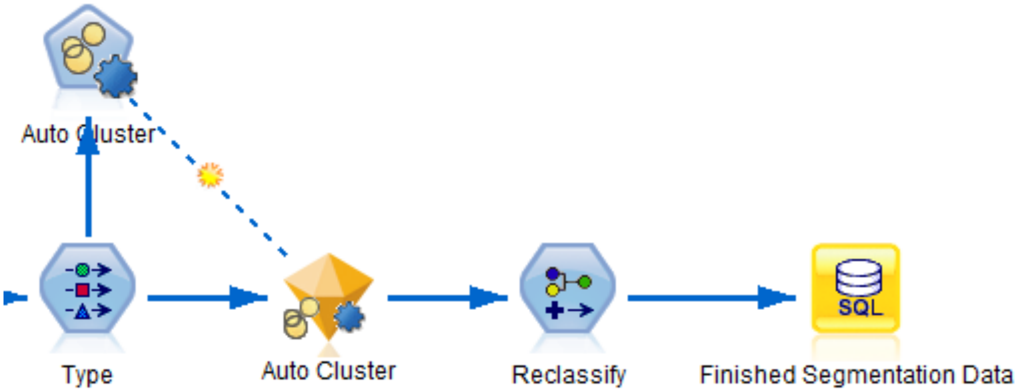


The resulting table now includes a new column with the cluster assignments.



	Income in Thousands	Number of Children	Level of Education_	Age Category_	Total Spend	Number Transactions	$XC-autocluster
1	140.000	3.000	Some High School	35 - 49	21.500	13.000	Low Value Married No Kids
2	140.000	3.000	Some High School	35 - 49	21.500	13.000	Low Value Married No Kids
3	225.000	4.000	High School Graduate	35 - 49	28.400	17.000	Average Married No Kids
4	225.000	4.000	High School Graduate	35 - 49	28.400	17.000	Average Married No Kids
5	225.000	3.000	High School Graduate	35 - 49	28.400	17.000	Average Married No Kids
6	225.000	3.000	High School Graduate	35 - 49	28.400	17.000	Average Married No Kids
7	210.000	4.000	High School Graduate	50 - 64	42.000	18.000	Average Married No Kids
8	210.000	4.000	High School Graduate	50 - 64	42.000	18.000	Average Married No Kids
9	150.000	3.000	Some High School	35 - 49	23.990	16.000	Low Value Married No Kids
10	150.000	3.000	Some High School	35 - 49	23.990	16.000	Low Value Married No Kids
11	200.000	4.000	High School Graduate	50 - 64	33.950	18.000	Average Married No Kids
12	200.000	4.000	High School Graduate	50 - 64	33.950	18.000	Average Married No Kids
13	310.000	4.000	College Graduate	50 - 64	62.000	23.000	High Income Families
14	310.000	4.000	College Graduate	50 - 64	62.000	23.000	High Income Families
15	170.000	3.000	High School Graduate	35 - 49	26.990	16.000	Average Married No Kids
16	170.000	3.000	High School Graduate	35 - 49	26.990	16.000	Average Married No Kids
17	193.000	3.000	High School Graduate	50 - 64	33.400	16.000	Average Married No Kids
18	193.000	3.000	High School Graduate	50 - 64	33.400	16.000	Average Married No Kids
19	193.000	3.000	High School Graduate	50 - 64	38.900	18.000	Average Married No Kids
20	193.000	3.000	High School Graduate	50 - 64	38.900	18.000	Average Married No Kids

You can also export these results back into the original data set or into other formats for use in later analyses. The following stream illustrates this, though we will not construct it here. Instead, for the purposes of this workshop, this step was already done and the data exported using the SQL Export node (shown below).



Summary

- ✓ Merge disparate data sources, including customer data from a database or CRM
- ✓ Define which fields to use
- ✓ Automatically generate a model to group customers
- ✓ Apply business terms to grouped customers
- ✓ Send new groups to database

For this exercise, we merged two data sources together, one from a database, and the other from a locally stored flat file. In order to identify groupings within our data, we used an automated clustering technique, specifying desired parameters. The resulting clusters were reclassified into business terms and exported back to the database.

Exercise 3: Understand the Past, Predict the Future

Use Case

Goal: Identify who is likely to respond to a marketing offer
Approach:

- Use a data extract from a CRM
- Extract concepts from the open ended comments in a customer survey
- Define which fields to use
- Choose the modeling technique
- Automatically generate a model to identify who is likely to respond
- Review results

Why?

- Target those likely to respond to offers to increase revenue, cut costs
- Using unstructured data improves modeling accuracy and provides more insight

Customer Reference

A wireless communications provider in the United States uses predictive modeling of customer data to increase revenue by billions and reduce its customer churn rate to less than 1 percent, lower than any of its competitors.

Business challenge
Reactive and reflective marketing strategy is giving way to predictive modeling. One wireless communications company in the United States knew customer churn was hitting its bottom line and began looking for a solution that would enable it to deepen its customer focus by proactively targeting customers more prone to churn.

The transformation
By mining both structured and unstructured customer data more successfully and feeding that data into more than 40 types of predictive models, the company could more accurately predict factors such as satisfaction levels, upgrade paths and the next best action to take to benefit both customers and the company.

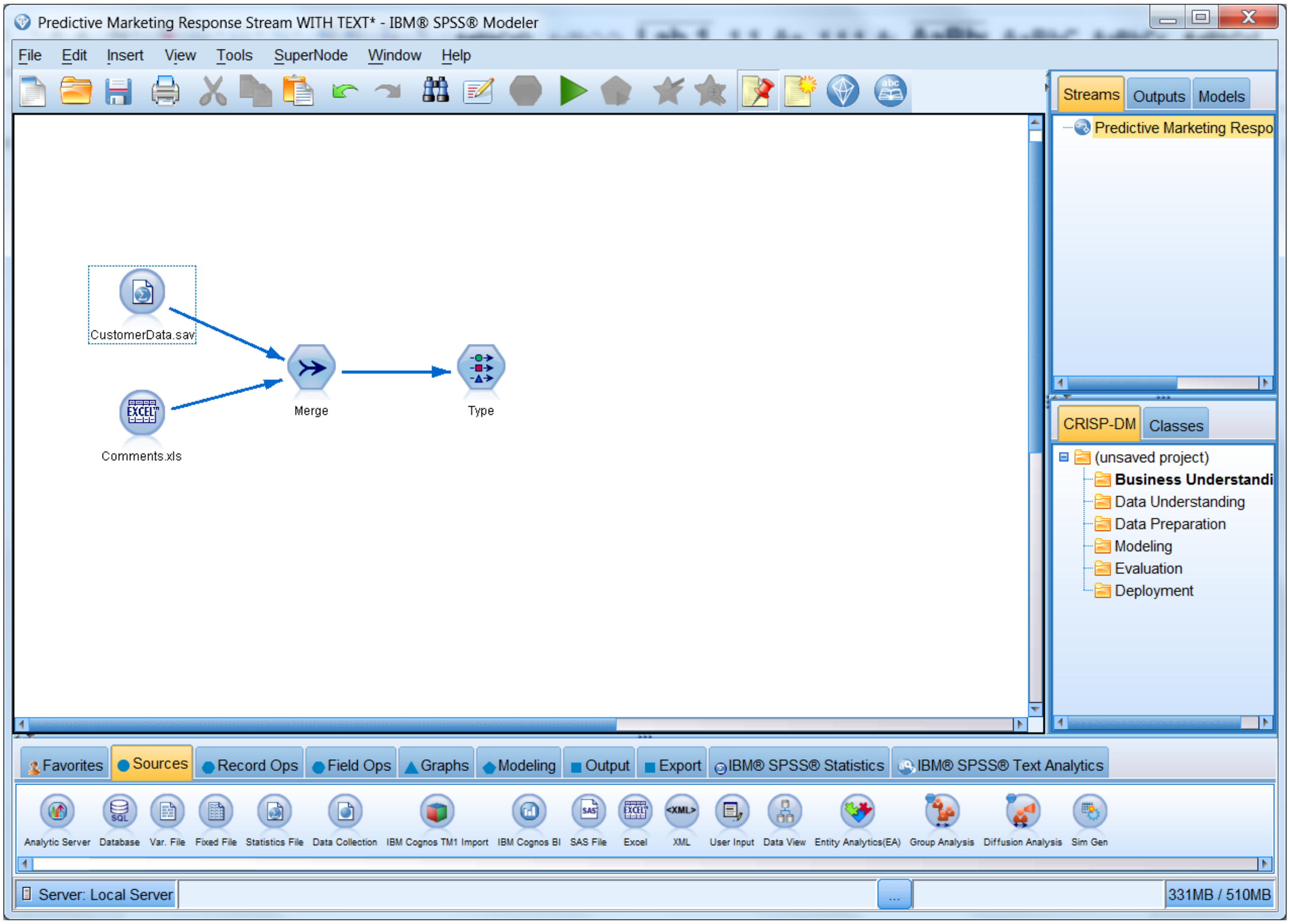
IBM's implemented solution:
Reduced customer churn by two-thirds to 0.94 percent, the lowest churn of any wireless provider in the country.
Grew company revenue by 7 billion in one year, a 10 percent increase.
Increased modeling accuracy for data by as much as 12 percent.
Enabled the company to evaluate more than 450 variables for predicting customer defection within 90 days.

"Enhanced predictive modeling not only helps us retain valued customers, but also helps us do it in a way that best preserves and enhances company profitability and alignment with our business goals."

Understand the Past, Predict the Future

1. Open the “Classifcation Exercise.str” file from the workshop directory. In IBM SPSS Modeler, click on File, Open stream, and then navigate to C:\Modeler Workshop\Classification\Classification Exercise.str.

Either double-click on Classification Exercise.str, or select it and then click on Open.

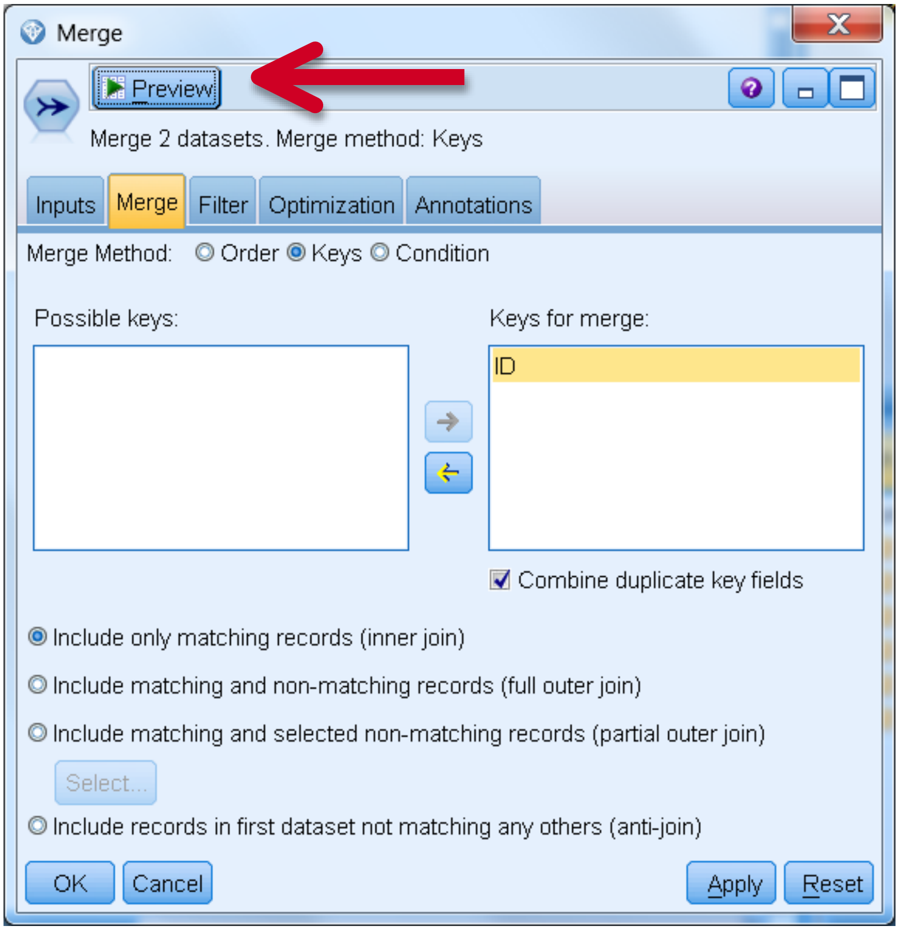


This excercise begins with a stream that has already been partially constructed. Note, in particular, that we are using the same customer data from the Predictive in 20 Minutes exercise. Notice also that we will be using unstructured data in the form of customer comments, and contained in a Microsoft Excel file.

2. From the Record Ops palette, a Merge node has already been placed on the canvas.

Double-click on the Merge node to review the settings. You will notice that the two data sources on the canvas are being joined by a common Key, in this case "ID".

Select Preview to see the first 10 records of this new merge.



Merge 2 datasets. Merge method: Keys

Merge Method: ○ Order ● Keys ○ Condition

Possible keys:

Keys for merge:
ID

☑ Combine duplicate key fields

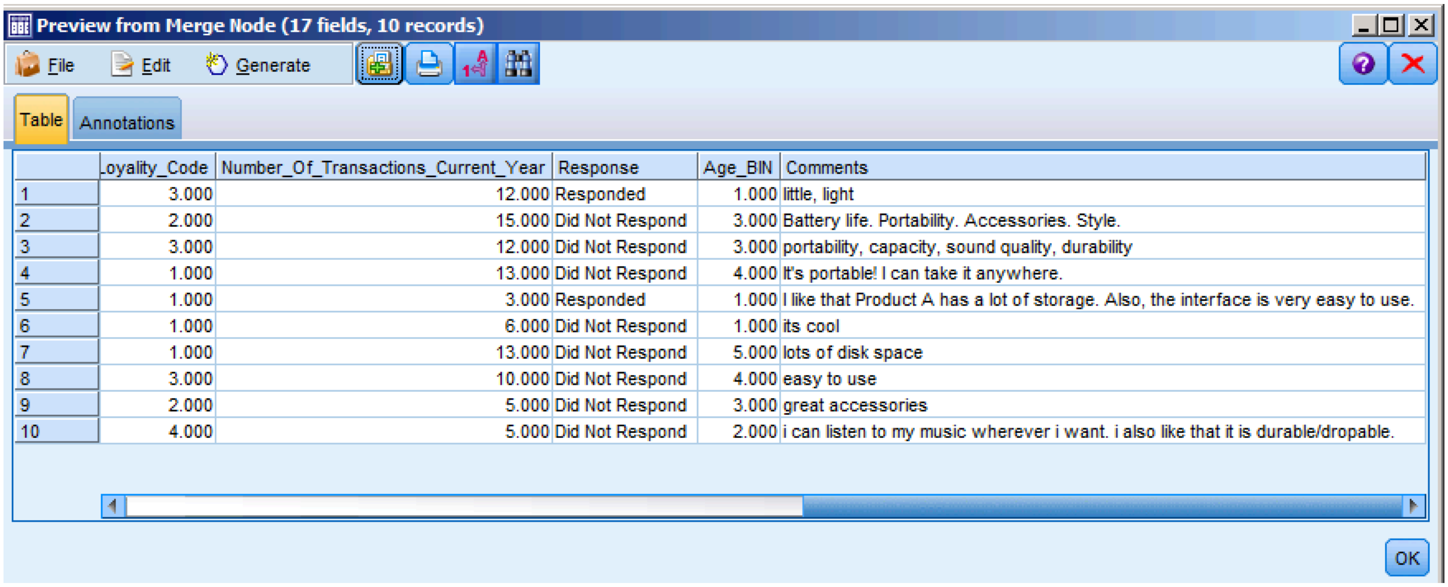
● Include only matching records (inner join)

○ Include matching and non-matching records (full outer join)

○ Include matching and selected non-matching records (partial outer join)

○ Include records in first dataset not matching any others (anti-join)

3. Scrolling to the right, you will notice the assigned cluster for each customer, created in the previous exercise; and a newly merged column, containing unstructured customer comments.

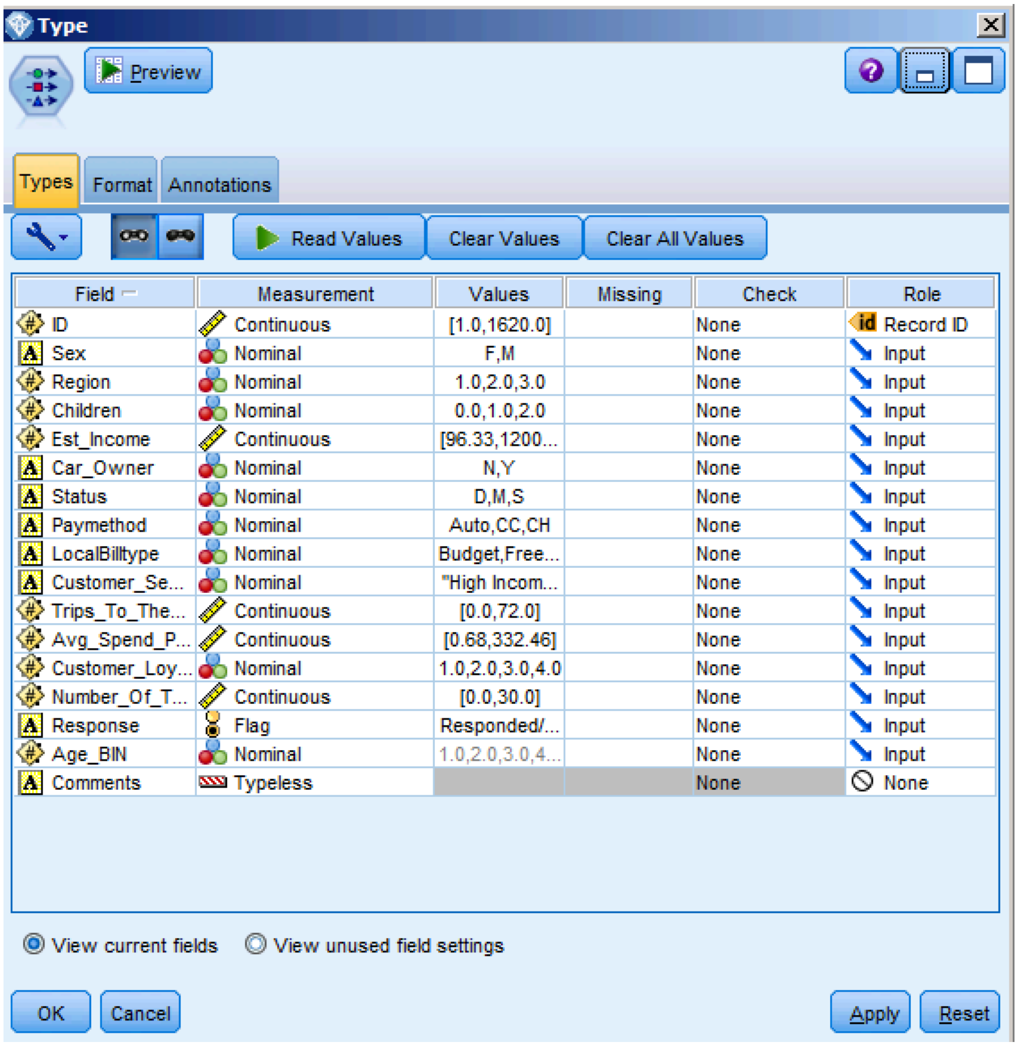


Preview from Merge Node (17 fields, 10 records)

	Loyality_Code	Number_Of_Transactions_Current_Year	Response	Age_BIN	Comments
1	3.000	12.000	Responded	1.000	little, light
2	2.000	15.000	Did Not Respond	3.000	Battery life. Portability. Accessories. Style.
3	3.000	12.000	Did Not Respond	3.000	portability, capacity, sound quality, durability
4	1.000	13.000	Did Not Respond	4.000	It's portable! I can take it anywhere.
5	1.000	3.000	Responded	1.000	I like that Product A has a lot of storage. Also, the interface is very easy to use.
6	1.000	6.000	Did Not Respond	1.000	its cool
7	1.000	13.000	Did Not Respond	5.000	lots of disk space
8	3.000	10.000	Did Not Respond	4.000	easy to use
9	2.000	5.000	Did Not Respond	3.000	great accessories
10	4.000	5.000	Did Not Respond	2.000	i can listen to my music wherever i want. i also like that it is durable/dropable.

4. After the merge, a Type node has been added to read the data and define the measurement level and role of each Field in the analysis.

   Double-click on the Type node to review its settings. An additional input, Comments, has been added as a result of the merge.



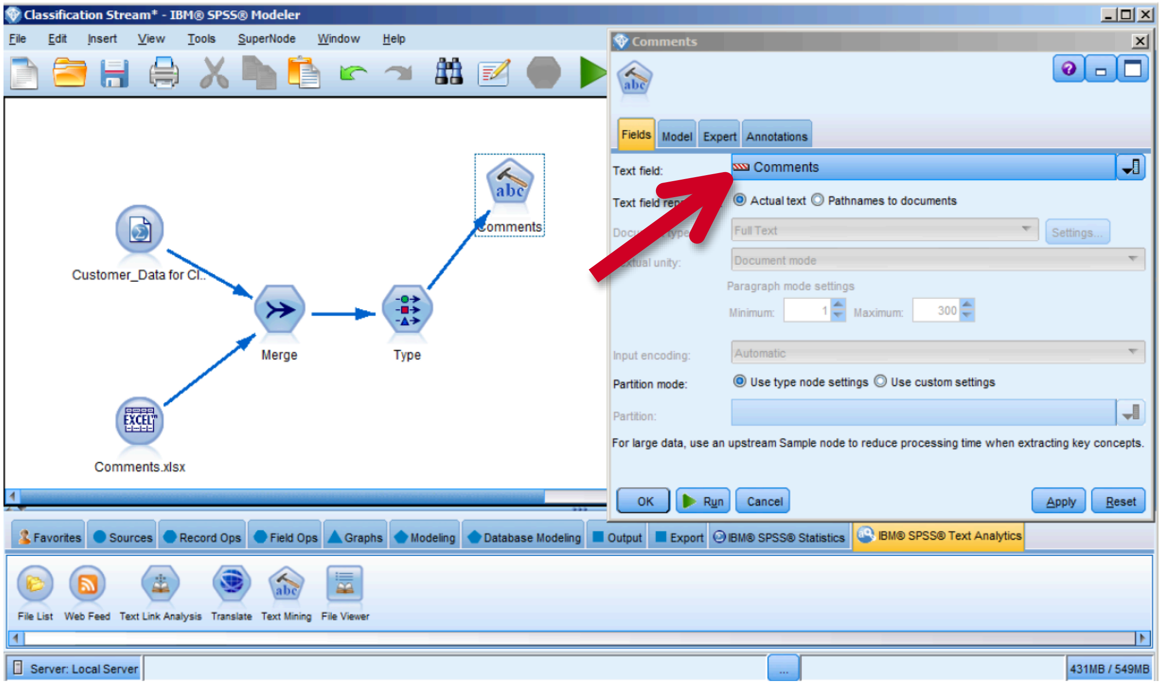
Field	Measurement	Values	Missing	Check	Role
ID	Continuous	[1.0,1620.0]		None	Record ID
Sex	Nominal	F,M		None	Input
Region	Nominal	1.0,2.0,3.0		None	Input
Children	Nominal	0.0,1.0,2.0		None	Input
Est_Income	Continuous	[96.33,1200...		None	Input
Car_Owner	Nominal	N,Y		None	Input
Status	Nominal	D,M,S		None	Input
Paymethod	Nominal	Auto,CC,CH		None	Input
LocalBilltype	Nominal	Budget,Free...		None	Input
Customer_Se...	Nominal	"High Incom...		None	Input
Trips_To_The...	Continuous	[0.0,72.0]		None	Input
Avg_Spend_P...	Continuous	[0.68,332.46]		None	Input
Customer_Loy...	Nominal	1.0,2.0,3.0,4.0		None	Input
Number_Of_T...	Continuous	[0.0,30.0]		None	Input
Response	Flag	Responded/...		None	Input
Age_BIN	Nominal	1.0,2.0,3.0,4...		None	Input
Comments	Typeless			None	None

● View current fields ○ View unused field settings

Using Text Analytics, we can identify patterns in the unstructured data and from them create categories which contain the *ideas* and *sentiments* as expressed by customers.

5. From the IBM SPSS Text Analytics palette, add a Text mining node to the canvas, and connect it to the Type node.

Double-click on the Text Mining node to review its settings. Using the Text field drop down, select Comments.


Classification Stream* - IBM® SPSS® Modeler
File Edit Insert View Tools SuperNode Window Help
Customer_Data for Cl...
Comments.xlsx
Merge
Type
Comments
Favorites Sources Record Ops Field Ops Graphs Modeling Database Modeling Output Export IBM® SPSS® Statistics IBM® SPSS® Text Analytics
File List Web Feed Text Link Analysis Translate Text Mining File Viewer
Server: Local Server
431MB / 549MB
Fields Model Expert Annotations
OK Run Cancel Apply Reset

Comments

Text field: Comments

Text field represents: ● Actual text ○ Pathnames to documents

Document type: Full Text

Textual unity: Document mode

Paragraph mode settings

Minimum: 1 Maximum: 300

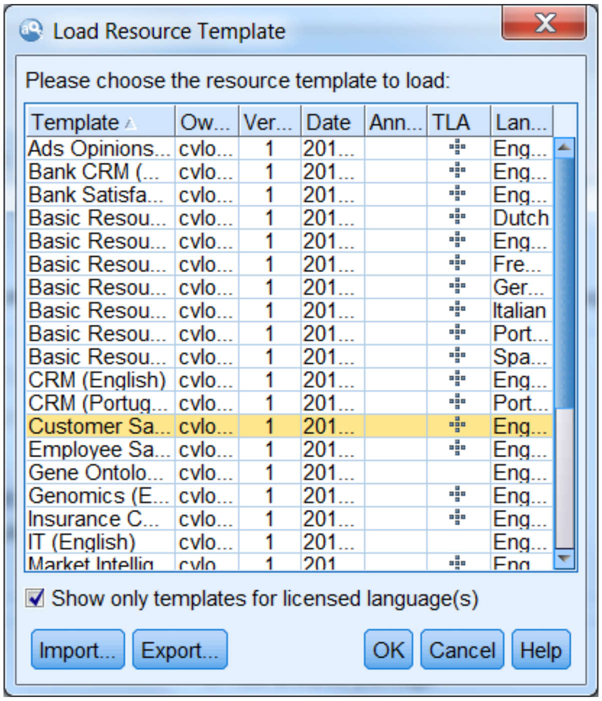
Input encoding: Automatic

Partition mode: ● Use type node settings ○ Use custom settings

Partition:

For large data, use an upstream Sample node to reduce processing time when extracting key concepts.

6. Select the Model tab and scroll down to the 'Copy Resources From' section. To select a Resource template, select Load and scroll to Customer Satisfaction Opinions (English) Library. This will load pre-built resources into the text mining process. Select OK and then Run.



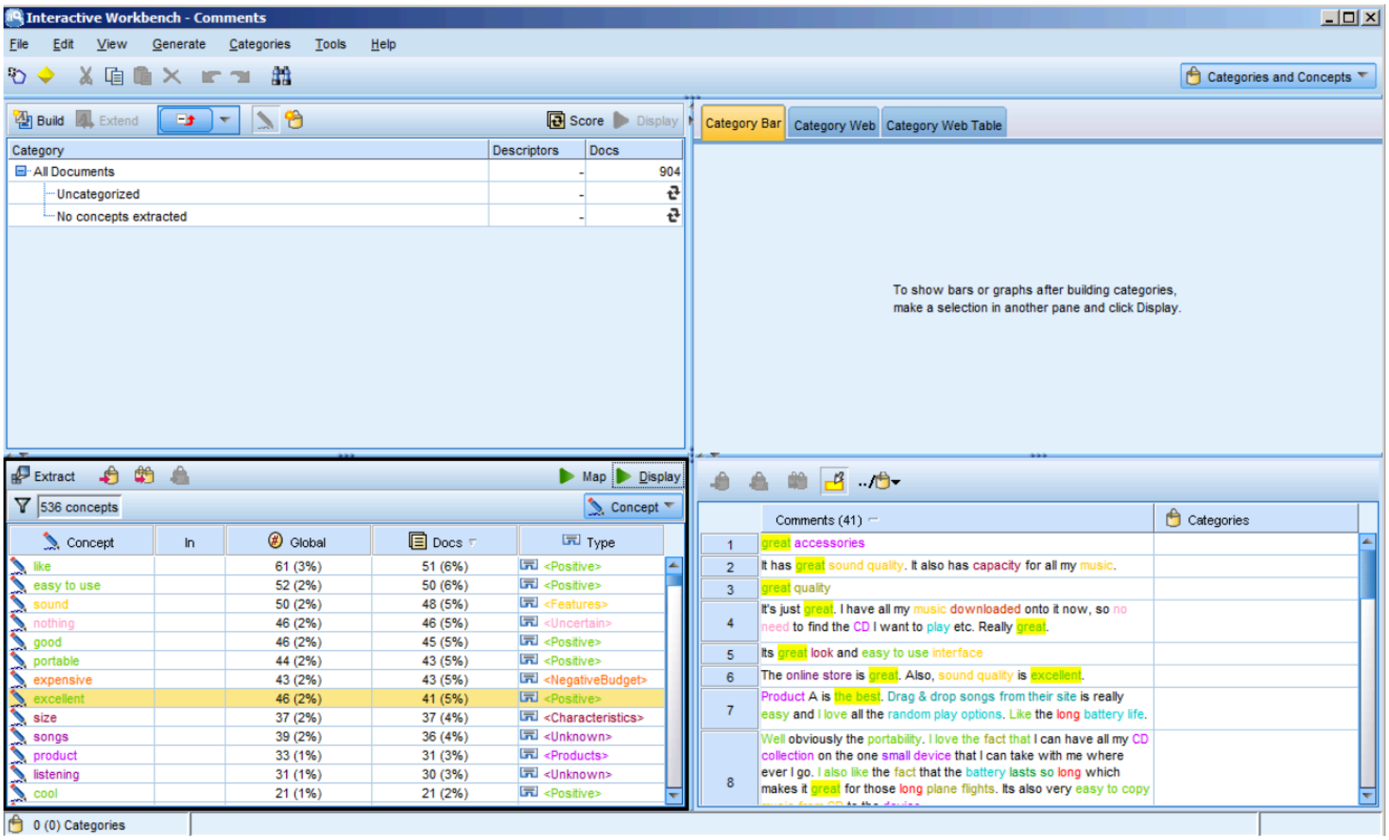
Load Resource Template

Please choose the resource template to load:

Template	Ow...	Ver...	Date	Ann...	TLA	Lan...
Ads Opinions...	cvlo...	1	201...		+	Eng...
Bank CRM (...	cvlo...	1	201...		+	Eng...
Bank Satisfa...	cvlo...	1	201...		+	Eng...
Basic Resou...	cvlo...	1	201...		+	Dutch
Basic Resou...	cvlo...	1	201...		+	Eng...
Basic Resou...	cvlo...	1	201...		+	Fre...
Basic Resou...	cvlo...	1	201...		+	Ger...
Basic Resou...	cvlo...	1	201...		+	Italian
Basic Resou...	cvlo...	1	201...		+	Port...
Basic Resou...	cvlo...	1	201...		+	Spa...
CRM (English)	cvlo...	1	201...		+	Eng...
CRM (Portug...	cvlo...	1	201...		+	Port...
Customer Sa...	cvlo...	1	201...		+	Eng...
Employee Sa...	cvlo...	1	201...		+	Eng...
Gene Ontolo...	cvlo...	1	201...			Eng...
Genomics (E...	cvlo...	1	201...		+	Eng...
Insurance C...	cvlo...	1	201...		+	Eng...
IT (English)	cvlo...	1	201...			Eng...
Market Intellig	cvlo	1	201		+	Eng

☑ Show only templates for licensed language(s)

7. Once the libraries and resources are loaded and the extraction process is complete, the Interactive Workbench is displayed. Note the list of extracted concepts in the lower left panel of the interface. These are not just words, phrases or character strings which were matched to some search criteria, but are *concepts* as identified, using Natural Language Processing (NLP), through reference to a comprehensive collection of libraries, provided with IBM SPSS Text Analytics for Modeler. They are those concepts on which our categorizations will be built. While in practice, Text Mining is a reiterative and interactive effort, for this workshop, we will run the text analysis engine without making any changes to the defaults.



Category	Descriptors	Docs
All Documents	-	904
Uncategorized	-	
No concepts extracted	-	

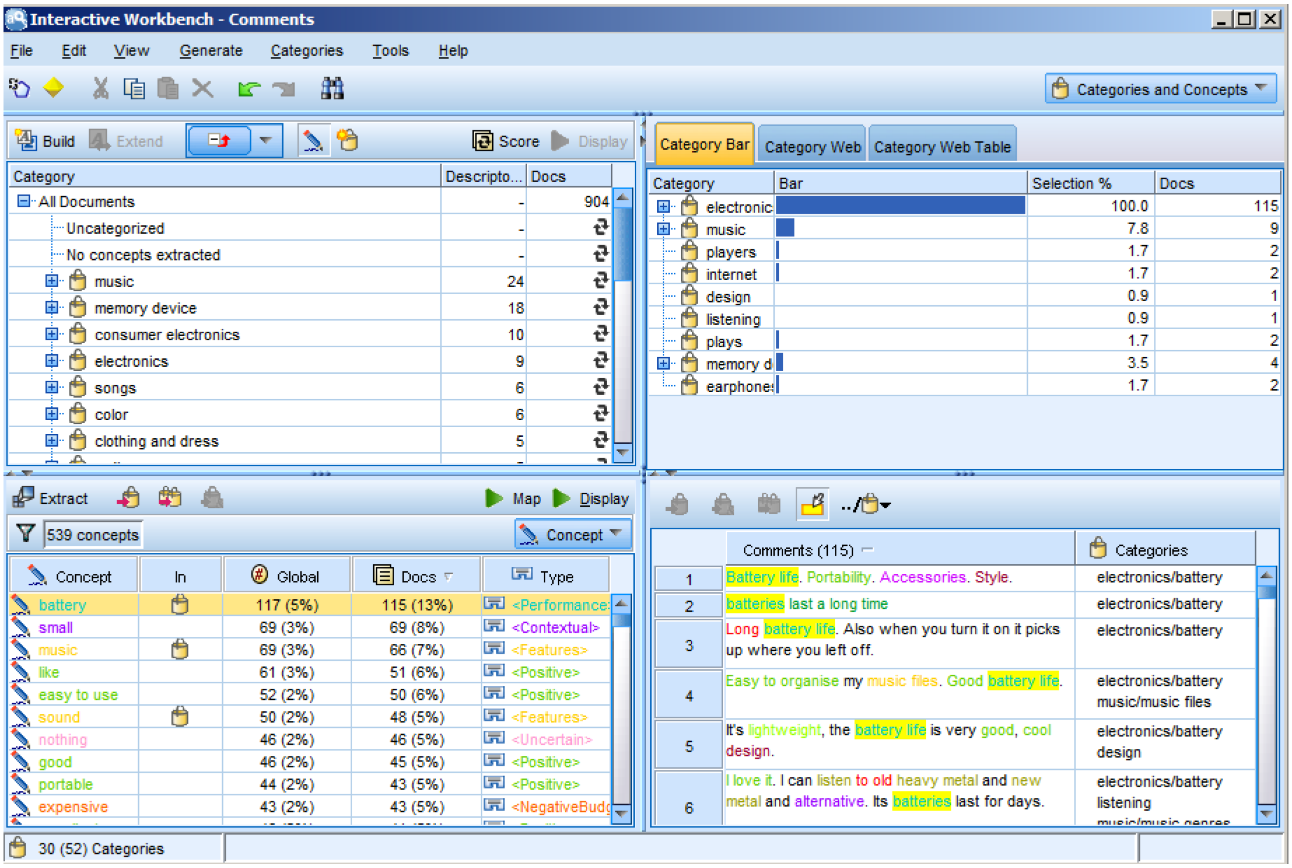
To show bars or graphs after building categories, make a selection in another pane and click Display.

Concept	In	Global	Docs	Type
like		61 (3%)	51 (6%)	<Positive>
easy to use		52 (2%)	50 (6%)	<Positive>
sound		50 (2%)	48 (5%)	<Features>
nothing		46 (2%)	46 (5%)	<Uncertain>
good		46 (2%)	45 (5%)	<Positive>
portable		44 (2%)	43 (5%)	<Positive>
expensive		43 (2%)	43 (5%)	<NegativeBudget>
excellent		46 (2%)	41 (5%)	<Positive>
size		37 (2%)	37 (4%)	<Characteristics>
songs		39 (2%)	36 (4%)	<Unknown>
product		33 (1%)	31 (3%)	<Products>
listening		31 (1%)	30 (3%)	<Unknown>
cool		21 (1%)	21 (2%)	<Positive>

	Comments (41)	Categories
1	great accessories	
2	It has great sound quality. It also has capacity for all my music.	
3	great quality	
4	It's just great. I have all my music downloaded onto it now, so no need to find the CD I want to play etc. Really great.	
5	Its great look and easy to use interface	
6	The online store is great. Also, sound quality is excellent.	
7	Product A is the best. Drag & drop songs from their site is really easy and I love all the random play options. Like the long battery life.	
8	Well obviously the portability. I love the fact that I can have all my CD collection on the one small device that I can take with me where ever I go. I also like the fact that the battery lasts so long which makes it great for those long plane flights. Its also very easy to copy	

8. To begin the process of building categories, click on Categories > Build Categories > Build now (not shown).

Once completed, it is at this step that the user would review the results, and then work with linguistic resources and category definitions to ultimately arrive at a set of categories, which are both useful and meaningful to the analysis. However, for purposes of this workshop, we will proceed with the categories as they are now. Click on the menu item for Generate > Generate Model. This will place the text mining model into the Models tray at the upper right corner of the workbench. Minimize the Interactive Workbench to return to your stream.



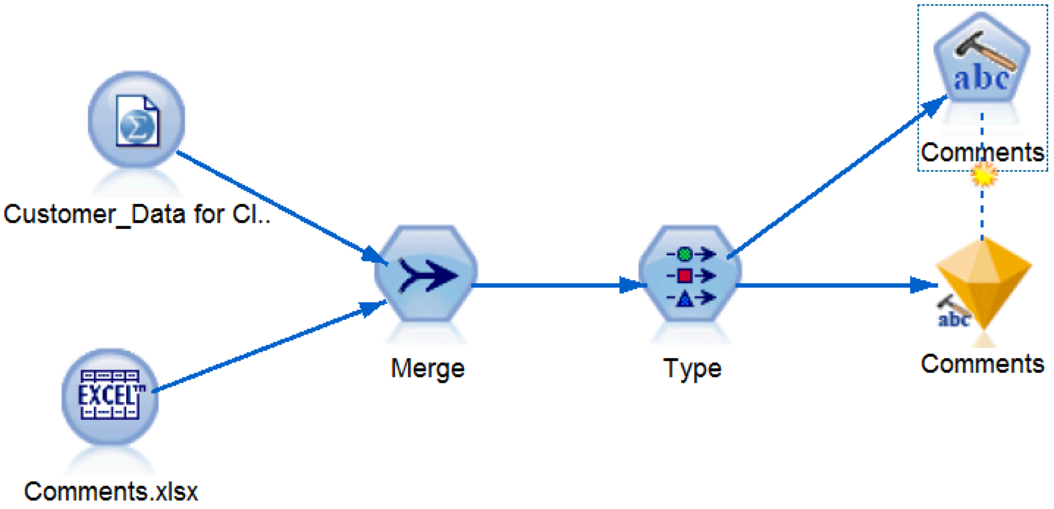
Category	Descripto...	Docs
All Documents	-	904
Uncategorized	-	
No concepts extracted	-	
music	24	
memory device	18	
consumer electronics	10	
electronics	9	
songs	6	
color	6	
clothing and dress	5	

Category	Bar	Selection %	Docs
electronic		100.0	115
music		7.8	9
players		1.7	2
internet		1.7	2
design		0.9	1
listening		0.9	1
plays		1.7	2
memory d		3.5	4
earphone:		1.7	2

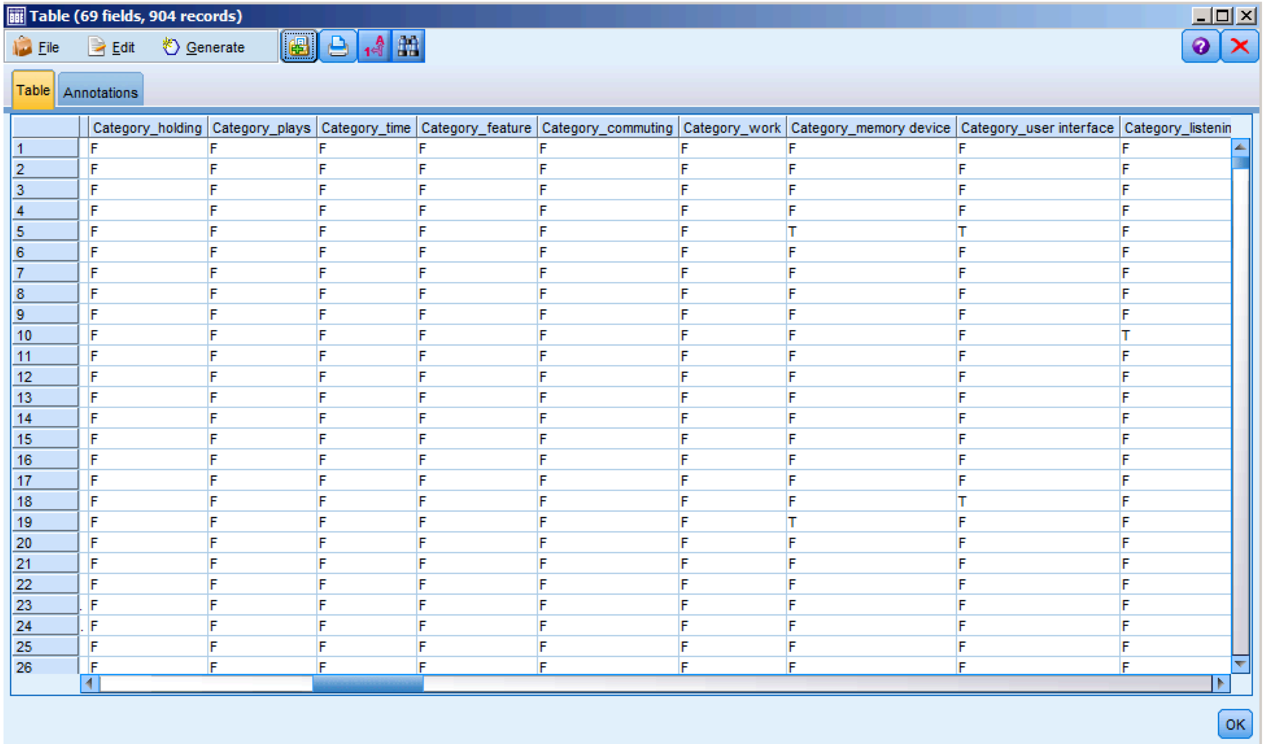
Concept	In	Global	Docs	Type
battery	✓	117 (5%)	115 (13%)	<Performance:
small		69 (3%)	69 (8%)	<Contextual>
music	✓	69 (3%)	66 (7%)	<Features>
like		61 (3%)	51 (6%)	<Positive>
easy to use		52 (2%)	50 (6%)	<Positive>
sound	✓	50 (2%)	48 (5%)	<Features>
nothing		46 (2%)	46 (5%)	<Uncertain>
good		46 (2%)	45 (5%)	<Positive>
portable		44 (2%)	43 (5%)	<Positive>
expensive		43 (2%)	43 (5%)	<NegativeBudg

	Comments (115)	Categories
1	Battery life. Portability. Accessories. Style.	electronics/battery
2	batteries last a long time	electronics/battery
3	Long battery life. Also when you turn it on it picks up where you left off.	electronics/battery
4	Easy to organise my music files. Good battery life.	electronics/battery music/music files
5	It's lightweight, the battery life is very good, cool design.	electronics/battery design
6	I love it. I can listen to old heavy metal and new metal and alternative. Its batteries last for days.	electronics/battery listening music/music genres

9. From the Models tray, drag the Comments model node onto the canvas, and connect it to the Type node.



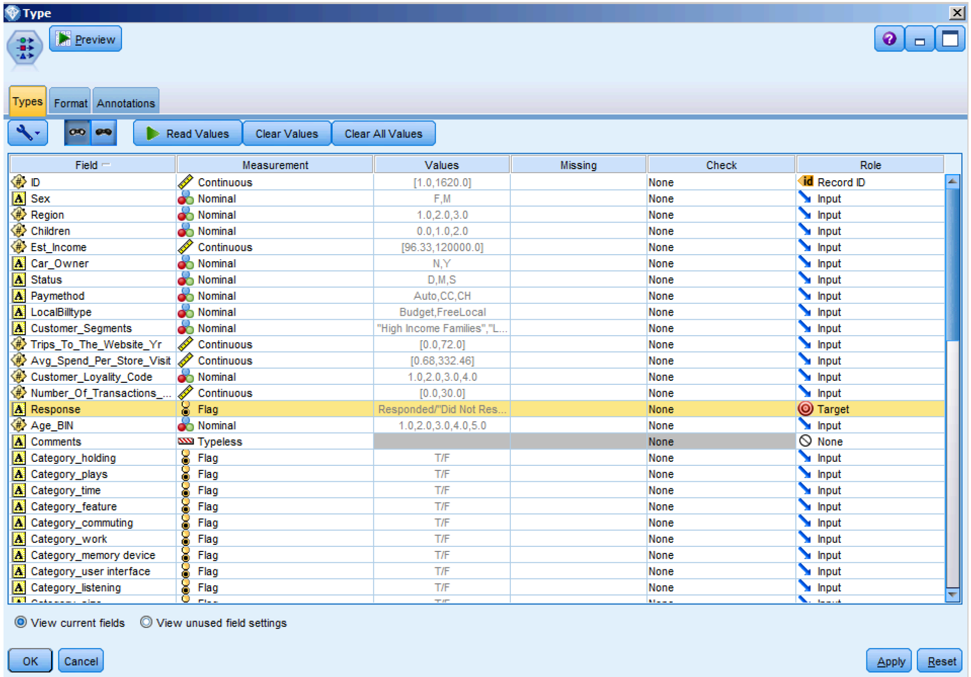
10. From the Output palette, connect a Table node to the modeling node and run it to see the comment categories. Scroll to the right to view the new categories created. There is a T if a comment was put into a category and an F it one was not.



	Category_holding	Category_plays	Category_time	Category_feature	Category_commuting	Category_work	Category_memory device	Category_user interface	Category_listenin
1	F	F	F	F	F	F	F	F	F
2	F	F	F	F	F	F	F	F	F
3	F	F	F	F	F	F	F	F	F
4	F	F	F	F	F	F	F	F	F
5	F	F	F	F	F	F	T	T	F
6	F	F	F	F	F	F	F	F	F
7	F	F	F	F	F	F	F	F	F
8	F	F	F	F	F	F	F	F	F
9	F	F	F	F	F	F	F	F	F
10	F	F	F	F	F	F	F	F	T
11	F	F	F	F	F	F	F	F	F
12	F	F	F	F	F	F	F	F	F
13	F	F	F	F	F	F	F	F	F
14	F	F	F	F	F	F	F	F	F
15	F	F	F	F	F	F	F	F	F
16	F	F	F	F	F	F	F	F	F
17	F	F	F	F	F	F	F	F	F
18	F	F	F	F	F	F	F	T	F
19	F	F	F	F	F	F	T	F	F
20	F	F	F	F	F	F	F	F	F
21	F	F	F	F	F	F	F	F	F
22	F	F	F	F	F	F	F	F	F
23	F	F	F	F	F	F	F	F	F
24	F	F	F	F	F	F	F	F	F
25	F	F	F	F	F	F	F	F	F
26	F	F	F	F	F	F	F	F	F

11. To properly define the newly created categories, we need to use a Type node. Add a Type node to the stream, connecting it to the modeling node.

Once connected, double-click on the Type node to open it. Click on Read Values to instantiate the data. Change the roles of ID to Record ID and Response to Target. Click OK. Notice, as you scroll down, the addition of many more inputs that are the result of structuring previously unstructured data.

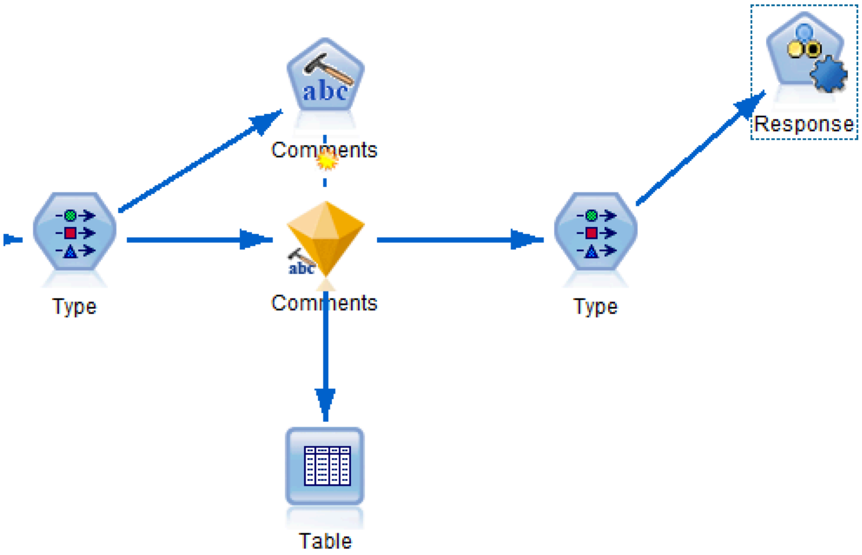


Field	Measurement	Values	Missing	Check	Role
ID	Continuous	[1.0,1620.0]		None	Record ID
Sex	Nominal	F,M		None	Input
Region	Nominal	1.0,2.0,3.0		None	Input
Children	Nominal	0.0,1.0,2.0		None	Input
Est_Income	Continuous	[96.33,120000.0]		None	Input
Car_Owner	Nominal	N,Y		None	Input
Status	Nominal	D,M,S		None	Input
Paymethod	Nominal	Auto,CC,CH		None	Input
LocalBilltype	Nominal	Budget,FreeLocal		None	Input
Customer_Segments	Nominal	"High Income Families","L...		None	Input
Trips_To_The_Website_Yr	Continuous	[0.0,72.0]		None	Input
Avg_Spend_Per_Store_Visit	Continuous	[0.68,332.46]		None	Input
Customer_Loyality_Code	Nominal	1.0,2.0,3.0,4.0		None	Input
Number_Of_Transactions_...	Continuous	[0.0,30.0]		None	Input
Response	Flag	Responded/"Did Not Res...		None	Target
Age_BIN	Nominal	1.0,2.0,3.0,4.0,5.0		None	Input
Comments	Typeless			None	None
Category_holding	Flag	T/F		None	Input
Category_plays	Flag	T/F		None	Input
Category_time	Flag	T/F		None	Input
Category_feature	Flag	T/F		None	Input
Category_commuting	Flag	T/F		None	Input
Category_work	Flag	T/F		None	Input
Category_memory device	Flag	T/F		None	Input
Category_user interface	Flag	T/F		None	Input
Category_listening	Flag	T/F		None	Input

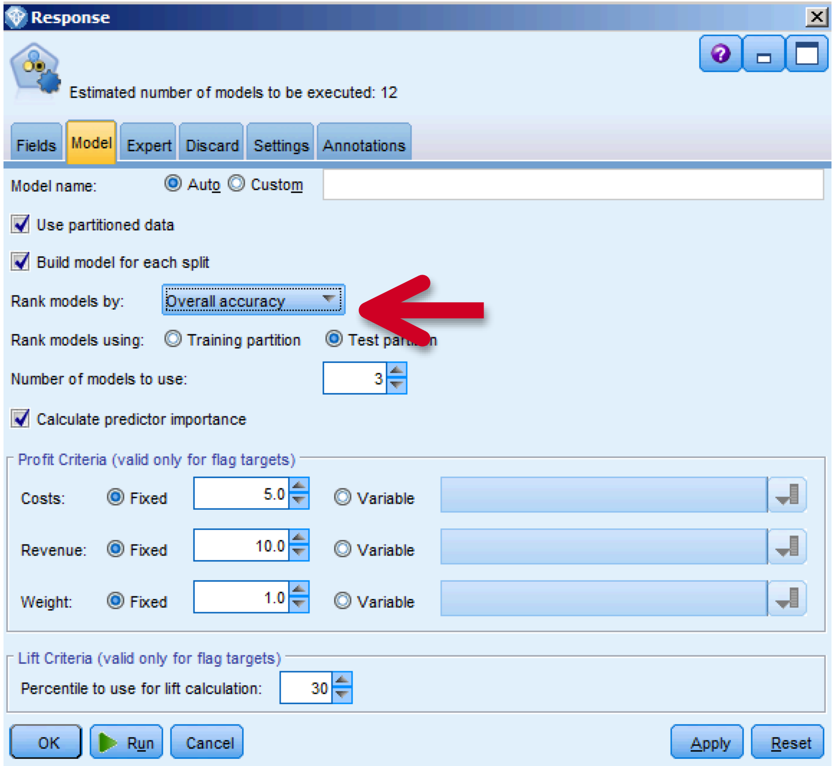
● View current fields ○ View unused field settings

12. To understand the impact these newly created Fields might have on our ability to predict campaign response, we will build another model.

Select the Auto Classifier node from the Modeling palette and connect it to the Type node. The Auto Classifier node creates and compares a number of different models for binary outcomes (yes or no, responded or did not respond, and so on), allowing you to choose the best approach for a given analysis.



Double-click on the Auto Classifier node to view the options. The Model tab of the Auto Classifier node enables you to specify the number of models to be created, along with the criteria used to compare models. For this exercise, we will be ranking models by Overall accuracy. The Expert tab (not shown) of the Auto Classifier node enables you to select the algorithms to use as well as model parameters. Click Run to execute your model.



Estimated number of models to be executed: 12

Model name: ● Auto ○ Custom

☑ Use partitioned data

☑ Build model for each split

Rank models by: Overall accuracy

Rank models using: ○ Training partition ● Test partition

Number of models to use: 3

☑ Calculate predictor importance

Profit Criteria (valid only for flag targets)

Costs: ● Fixed 5.0 ○ Variable

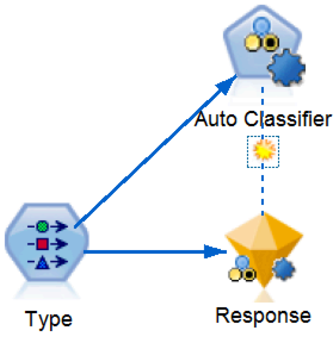
Revenue: ● Fixed 10.0 ○ Variable

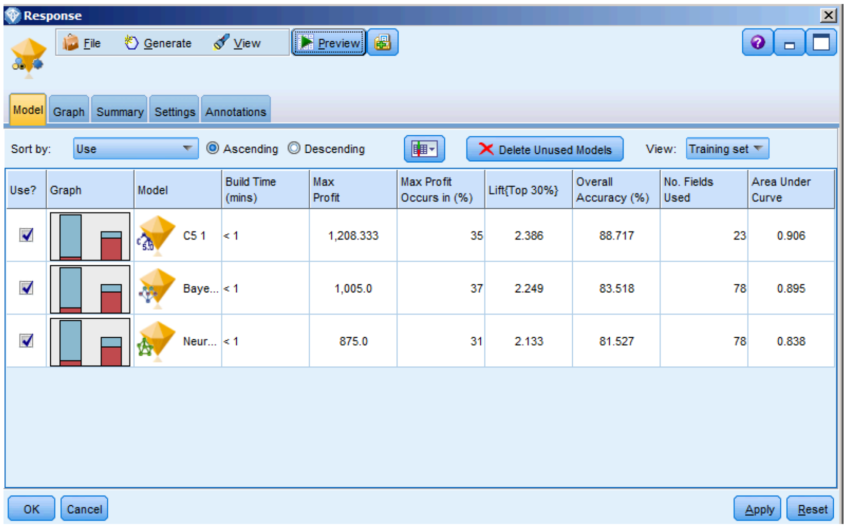
Weight: ● Fixed 1.0 ○ Variable

Lift Criteria (valid only for flag targets)

Percentile to use for lift calculation: 30

The Auto Classifier model is automatically generated and added to our canvas. When the Auto Classifier is executed, the node estimates candidate models for every possible combination of options, ranks each candidate model based on the measure you specify, and saves the best models in a composite automated model nugget. This model nugget contains a set of one or more models generated by the node, which can be individually browsed or selected for use in scoring.



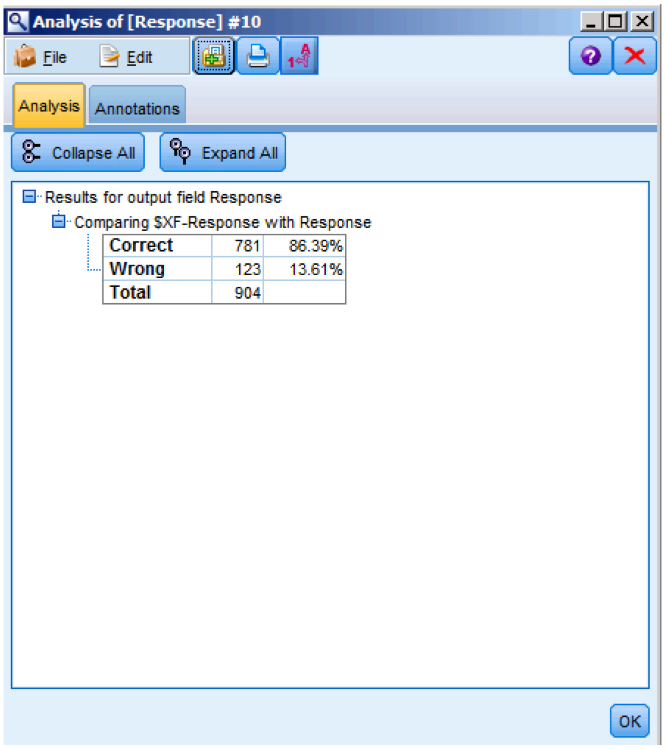


Sort by: Use ● Ascending ○ Descending View: Training set

Use?	Graph	Model	Build Time (mins)	Max Profit	Max Profit Occurs in (%)	Lift{Top 30%}	Overall Accuracy (%)	No. Fields Used	Area Under Curve
☑		C5 1	< 1	1,208.333	35	2.386	88.717	23	0.906
☑		Baye...	< 1	1,005.0	37	2.249	83.518	78	0.895
☑		Neur...	< 1	875.0	31	2.133	81.527	78	0.838

13. To view the overall accuracy of this model, from the Output palette, place and Analysis node on the canvas, connect it to the Response model, and select Run Selection from the toolbar.

The results from the Analysis node illustrates how, with the addition of unstructured data in the form of customer comments, we can increase the classification accuracy of the overall model. In this case, the accuracy is 86.39%. Recall that the accuracy of our Predictive in 20 Minutes model, which did not include unstructured data, was 84.25%.



Results for output field Response

Comparing $XF-Response with Response

Correct	781	86.39%
Wrong	123	13.61%
Total	904	

Summary

- ✓ Use a data extract from a CRM
- ✓ Extract concepts from the open ended comments in a customer survey
- ✓ Define which fields to use
- ✓ Choose the modeling technique
- ✓ Automatically generate a model to identify who is likely to respond
- ✓ Review results

In order to improve on the insights we have discovered thus far, related to understanding which customers are likely to respond to a campaign, we added available unstructured data in the form of customer comments. This resulted in our ability to understand sentiment related to our customers' experiences, improving the overall accuracy of our predictive model to identify those customers most likely to respond to future campaigns.

Exercise 3: Deployment

Use Case

Goal: Use trained models to score new customer data.
Approach:

- Use new customer records who have never received an offer
- Leverage the text extraction and classification models used in exercise 3
- Automatically generate scores of who is likely to respond
- Review results
- Deploy results for use by marketing team

Why?

- Build a targeted list of customers likely to respond to a campaign.

Customer Reference

A debt-collection firm identified debtors with the highest probability of repaying their creditors, thereby achieving 6-figure financial benefits when it implements an IBM predictive analytics solution.

Business Challenge
To keep its debt-recovery services profitable, a debt-collection firm needed to predict which individuals are most likely to settle their debts, devoting more calls, letters and research to sure bets rather than deadbeats.

The Transformation
The solution analyzes personal information, credit scores and census data*,* including average income, house value and level of education in the debtor's ZIP code, to predict each debtor's probability of repayment. The solution has transformed that time-consuming manual process into a fast, automated process that uses text analytics to mine closed files for key indicators of at-fault drivers, flagging top candidates for further review

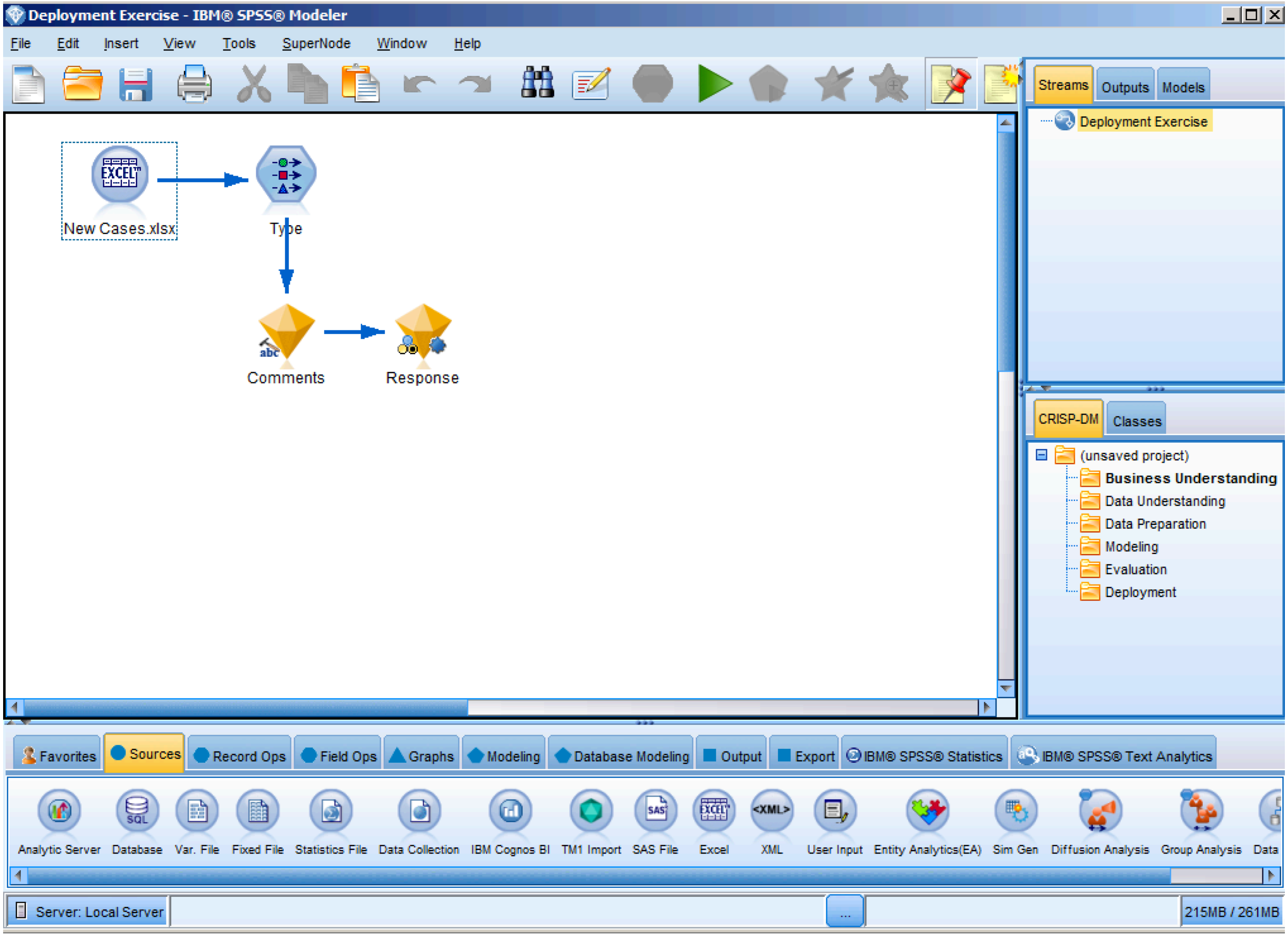
IBM's implemented solution resulted in:
8 million in potential cost savings by focusing on debtors most likely to pay and avoiding futile phone calls and emails
Six-figure savings as a result of eliminating outsourced analytics services and increasing collection success rates
94% reduction in manual processes, allowing the company to offer services at a more competitive price

"We had not anticipated just how accurate a predictor ZIP code information could be. Analysis has revealed that neighborhoods with higher-than-average incomes and house values are strong indicators of good payers—enabling us to focus our efforts on cases with the highest probability of returns."

Deployment

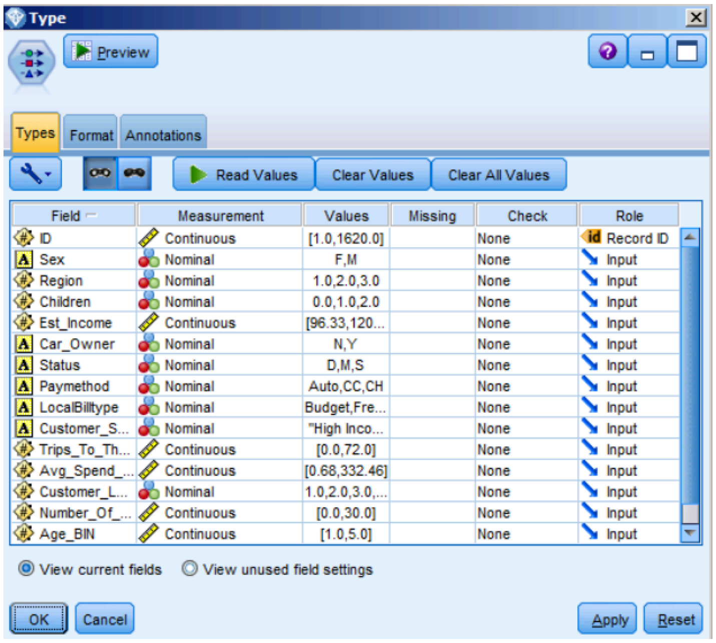
1. To start, open the stream labeled “Deployment Exercise.str” from the workshop directory. In IBM SPSS Modeler, click on File, Open stream, and then navigate to C:\Modeler Workshop\Deployment\Deployment Exercise.str.

Either double-click on “Deployment Exercise.str”, or select it and then click Open.



For this exericse, we are going to deploy the insights that we have discovered today against brand new customer data in order to score those records. This partially constructed stream inlcudes a new customer data file as well as our comments and prediction models.

Double-click on the Type node. Note that, because this is current customer data, we do not have a Response Field and no Fields have been set to Target.

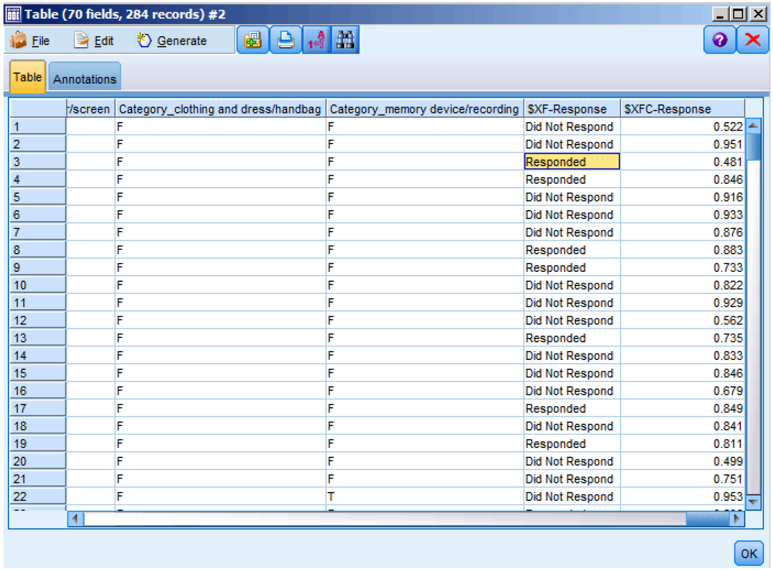


Field	Measurement	Values	Missing	Check	Role
ID	Continuous	[1.0,1620.0]		None	Record ID
Sex	Nominal	F,M		None	Input
Region	Nominal	1.0,2.0,3.0		None	Input
Children	Nominal	0.0,1.0,2.0		None	Input
Est_Income	Continuous	[96.33,120...		None	Input
Car_Owner	Nominal	N,Y		None	Input
Status	Nominal	D,M,S		None	Input
Paymethod	Nominal	Auto,CC,CH		None	Input
LocalBilltype	Nominal	Budget,Fre...		None	Input
Customer_S...	Nominal	"High Inco...		None	Input
Trips_To_Th...	Continuous	[0.0,72.0]		None	Input
Avg_Spend_...	Continuous	[0.68,332.46]		None	Input
Customer_L...	Nominal	1.0,2.0,3.0,...		None	Input
Number_Of_...	Continuous	[0.0,30.0]		None	Input
Age_BIN	Continuous	[1.0,5.0]		None	Input

● View current fields ○ View unused field settings

2. From the Output palette, add a Table node to the canvas, connect it to the Response model node, and select Run.

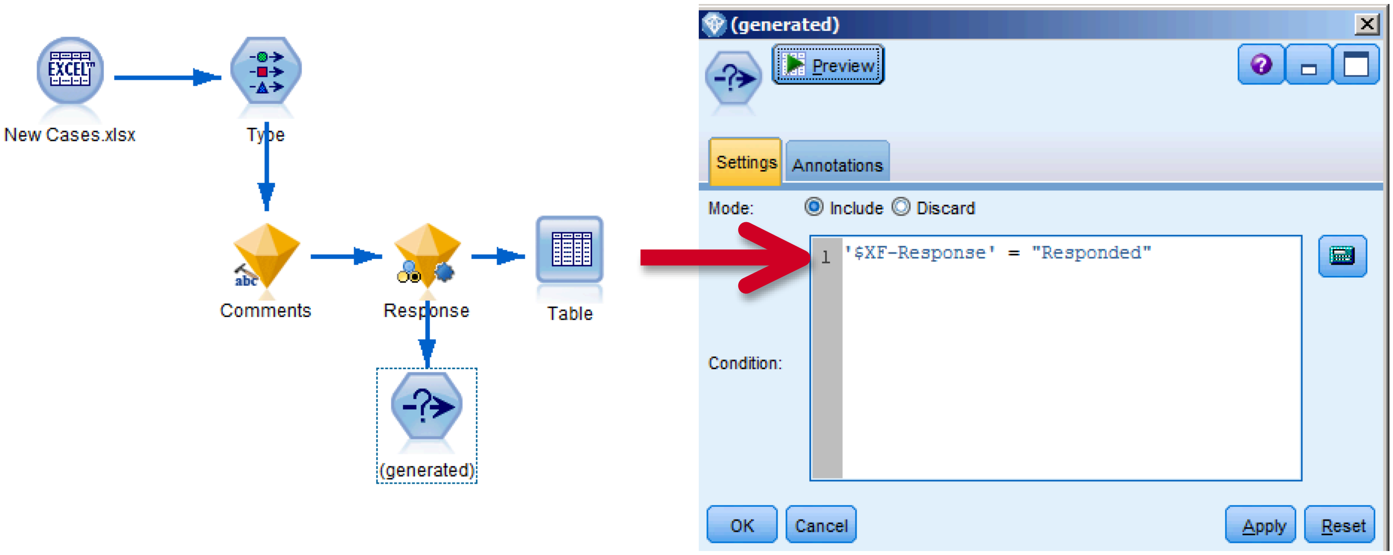
The new cases have been passed through the comments model to extract concepts, which were used as inputs in the Auto Classifier model to predict Response. The last two columns show the predicted outcome and the calculated confidence. For example, the first record in the table shown below is predicted not to respond with 52.2% confidence.



	r/screen	Category_clothing and dress/handbag	Category_memory device/recording	$XF-Response	$XFC-Response
1		F	F	Did Not Respond	0.522
2		F	F	Did Not Respond	0.951
3		F	F	Responded	0.481
4		F	F	Responded	0.846
5		F	F	Did Not Respond	0.916
6		F	F	Did Not Respond	0.933
7		F	F	Did Not Respond	0.876
8		F	F	Responded	0.883
9		F	F	Responded	0.733
10		F	F	Did Not Respond	0.822
11		F	F	Did Not Respond	0.929
12		F	F	Did Not Respond	0.562
13		F	F	Responded	0.735
14		F	F	Did Not Respond	0.833
15		F	F	Did Not Respond	0.846
16		F	F	Did Not Respond	0.679
17		F	F	Responded	0.849
18		F	F	Did Not Respond	0.841
19		F	F	Responded	0.811
20		F	F	Did Not Respond	0.499
21		F	F	Did Not Respond	0.751
22		F	T	Did Not Respond	0.953

3. Highlight one of the cells with the value of Responded. From the drop-down menu, select Generate from the tool bar and then choose Select Node (“Or”). This will generate a Select node on canvas. Alternatively, you can add a Select node form the Record Ops palette (not shown), and connect it to the Response model node.

Join the Generated node to the Response model node. Double-click on the node to review the settings, where only those records predicted with a Responded prediction are selected. Click OK.



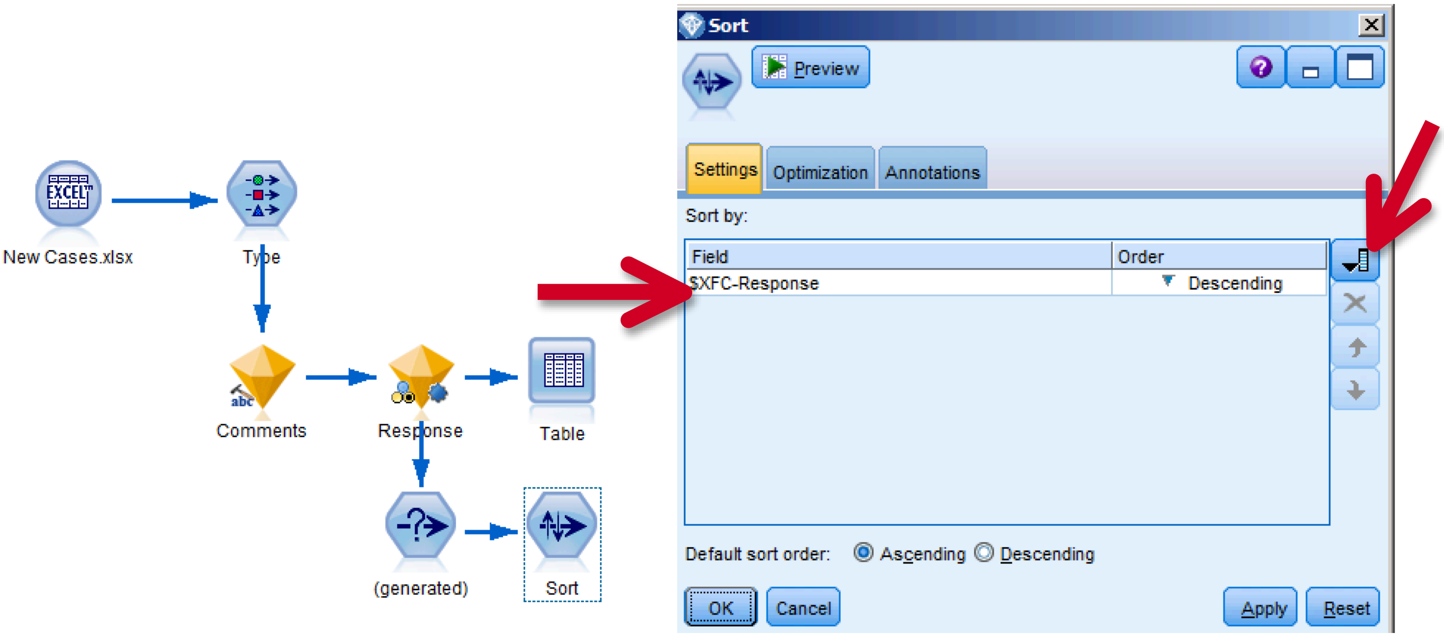
Mode: ● Include ○ Discard

Condition:

```
'$XF-Response' = "Responded"
```

4. From the Record Ops palette, select a Sort node, and connect it to the Generated node.

   Double-click to edit the settings, sorting the confidence score field, $XC-Response, in descending order. Here we are sorting our customers predicted to respond by confidence in that prediction, from highest confidence to lowest. Click OK.



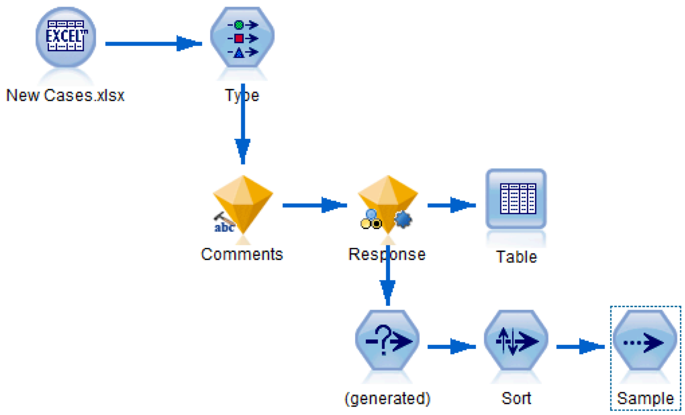
Sort by:

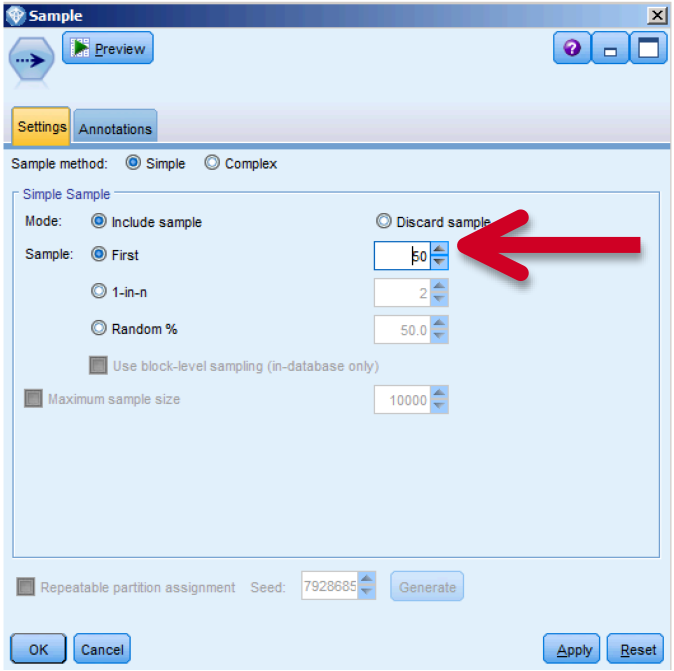
Field	Order
$XFC-Response	Descending

Default sort order: ● Ascending ○ Descending

5. Returning to the Record Ops palette, select a Sample node, and connect it to the Sort node.

   Double-click the Sample node to edit the settings, sampling the first 50 records. Click OK.





Sample method: ◉ Simple ○ Complex

Simple Sample

Mode: ◉ Include sample ○ Discard sample

Sample: ◉ First 50

○ 1-in-n 2

○ Random % 50.0

□ Use block-level sampling (in-database only)

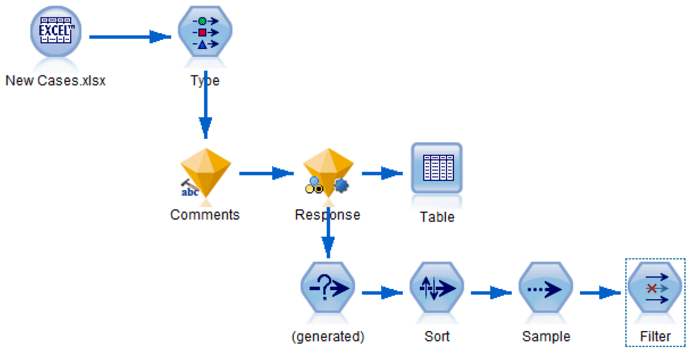
□ Maximum sample size 10000

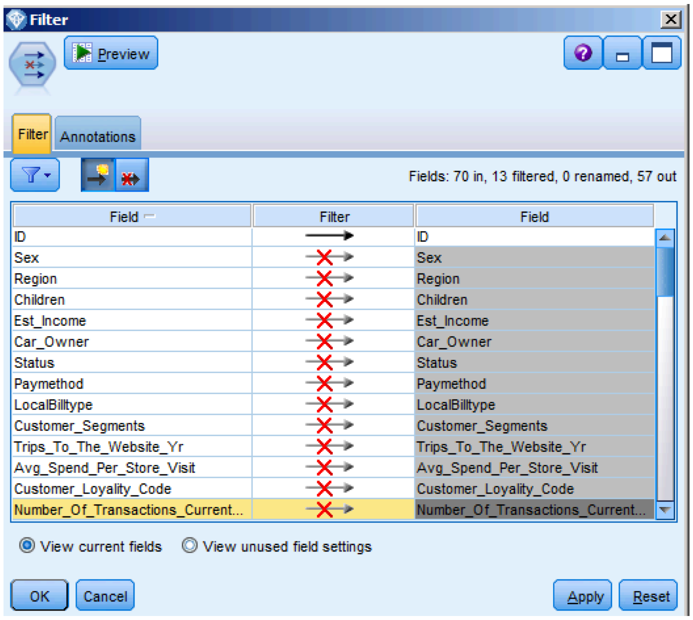
□ Repeatable partition assignment Seed: 7928685

6. Of interest to us now is a list of those customers predicted most likely to respond to a campaign. Having sorted and sampled our current customers, we can effectively create a list of the top 50 customers to target for a marketing campaign. That list can be filtered to include only relevant fields; and exported in a number of ways.

From the Fields Ops palette, select a Filter node, and connect it to the Sample node.

Double-click the Filter node to edit the settings, and Filter out Fields you do not want exported. In this case, we will filter all Fields except for ID, $XF- Response (predicted outcome), and $XFC- Response (confidence score).





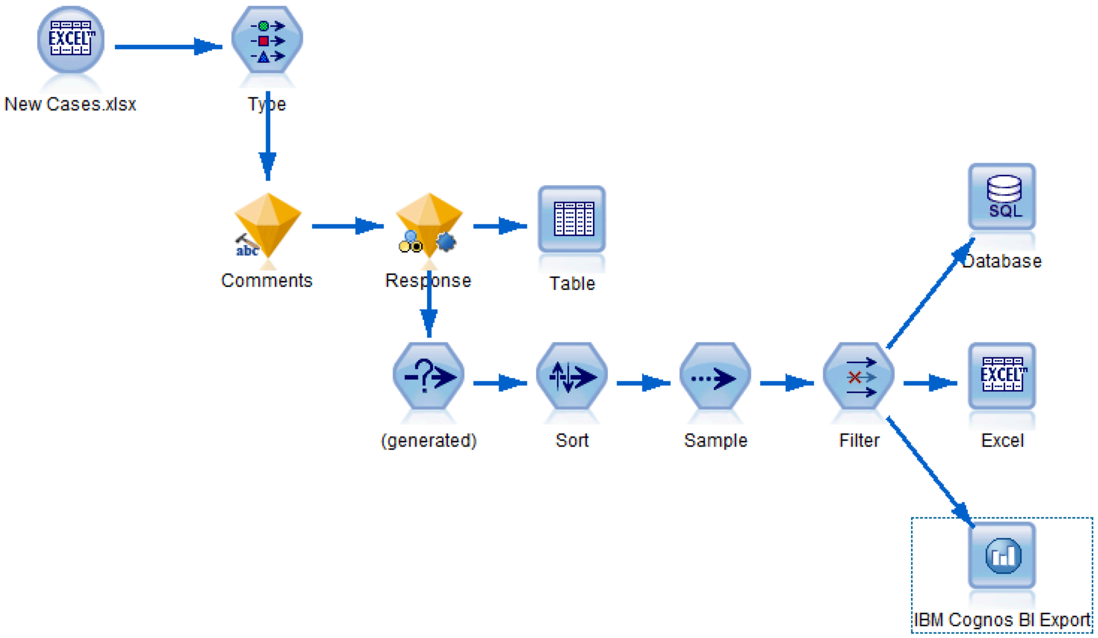
Fields: 70 in, 13 filtered, 0 renamed, 57 out

Field	Filter	Field
ID	→	ID
Sex	✗→	Sex
Region	✗→	Region
Children	✗→	Children
Est_Income	✗→	Est_Income
Car_Owner	✗→	Car_Owner
Status	✗→	Status
Paymethod	✗→	Paymethod
LocalBilltype	✗→	LocalBilltype
Customer_Segments	✗→	Customer_Segments
Trips_To_The_Website_Yr	✗→	Trips_To_The_Website_Yr
Avg_Spend_Per_Store_Visit	✗→	Avg_Spend_Per_Store_Visit
Customer_Loyality_Code	✗→	Customer_Loyality_Code
Number_Of_Transactions_Current...	✗→	Number_Of_Transactions_Current...

◉ View current fields ○ View unused field settings

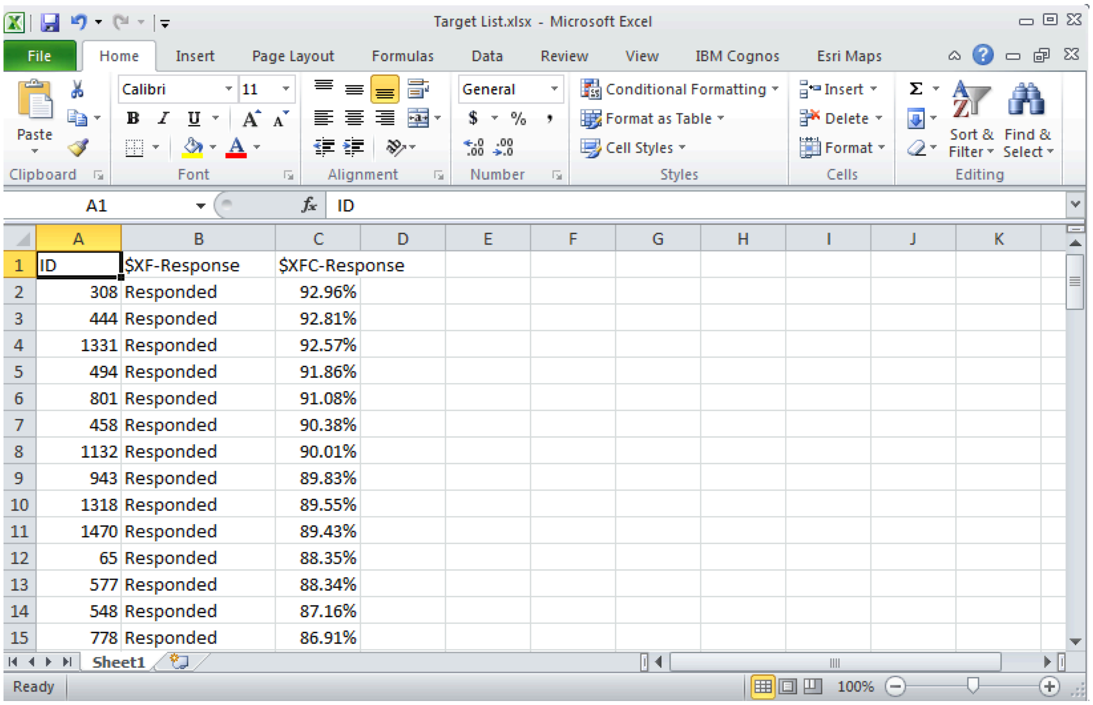
7. How to deploy these results depends on the use case. For example, marketing can be provided with a flat file, which can be merged with contact information. Raw scores can be exported to a database to maintain data integrity. Another option would be to export scores to IBM Cognos BI for inclusion in a C-Suite dashboard that will aid in decision-making.

   From the Export palette, select and connect the following to the Filter node: Excel, Database and IBM Cognos BI Export.



   Double-click on the Excel node to edit the settings. Change the File name to “Target List” in your local directory, and select “Launch Excel” before clicking Run.

   The result is actionable intelligence. That is, a list of the top 50 customer likely to respond to a marketing campaign.



ID	$XF-Response	$XFC-Response
308	Responded	92.96%
444	Responded	92.81%
1331	Responded	92.57%
494	Responded	91.86%
801	Responded	91.08%
458	Responded	90.38%
1132	Responded	90.01%
943	Responded	89.83%
1318	Responded	89.55%
1470	Responded	89.43%
65	Responded	88.35%
577	Responded	88.34%
548	Responded	87.16%
778	Responded	86.91%

Summary

- ✓ Use a data extract from a CRM
- ✓ Prepare data for modeling
- ✓ Define which fields to use
- ✓ Choose the modeling technique
- ✓ Automatically generate a model to identify who has responded
- ✓ Merge disparate data sources, including customer data from a database or CRM
- ✓ Automatically generate a model to group customers
- ✓ Apply business terms to new customer groups
- ✓ Export newly created groups to a database
- ✓ Extract concepts from the open ended comments in a customer survey
- ✓ Automatically generate a model to identify who is likely to respond
- ✓ Use new customer records who have never received an offer
- ✓ Leverage previously built text extraction and classification models
- ✓ Automatically generate scores of who is likely to respond
- ✓ Review results
- ✓ Deploy results for use by marketing team

For today's workshop, we built a predictive model to identify customers likely to respond to a campaign, segmented our customer base into like groups; and to improve model accuracy and leverage existing unstructured data, we built a text analysis model to capture concepts and sentiments in customer comments. Finally, those insights were deployed against new customer data, scoring those customers to identify those most likely to respond to a campaign. This allowed us to generate a list of customers to target for future campaigns, thereby reducing costs by increasing response rates.

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

THE INFORMATION IN THIS DOCUMENT IS PROVIDED "AS IS" WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT. IBM products are warranted according to the terms and conditions of the agreements under which they are provided.