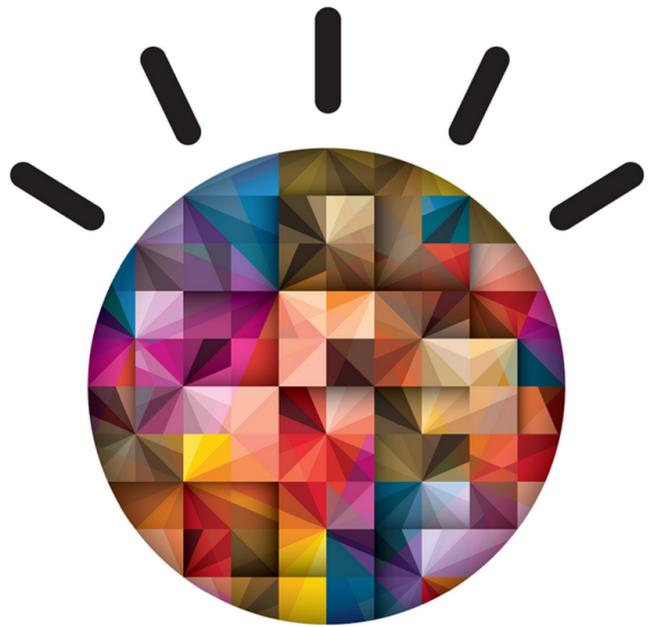


IBM Analytics



Data Science
Workshop Guide

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

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

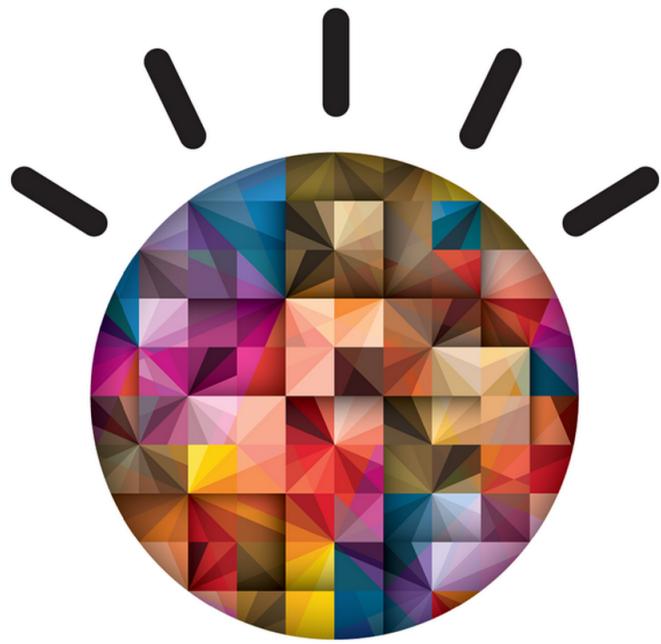
| | |
|---|-----------|
| Introduction..... | 3 |
| Exercise 1: Predictive in 20 Minutes | 4 |
| Exercise 2: Finding Patterns and Groups..... | 16 |
| Exercise 3: Understand the Past, Predict the Future..... | 30 |
| Exercise 4: Deployment..... | 42 |
| Summary | 48 |

Introduction

IBM SPSS Modeler is a comprehensive data science 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 Data Science 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

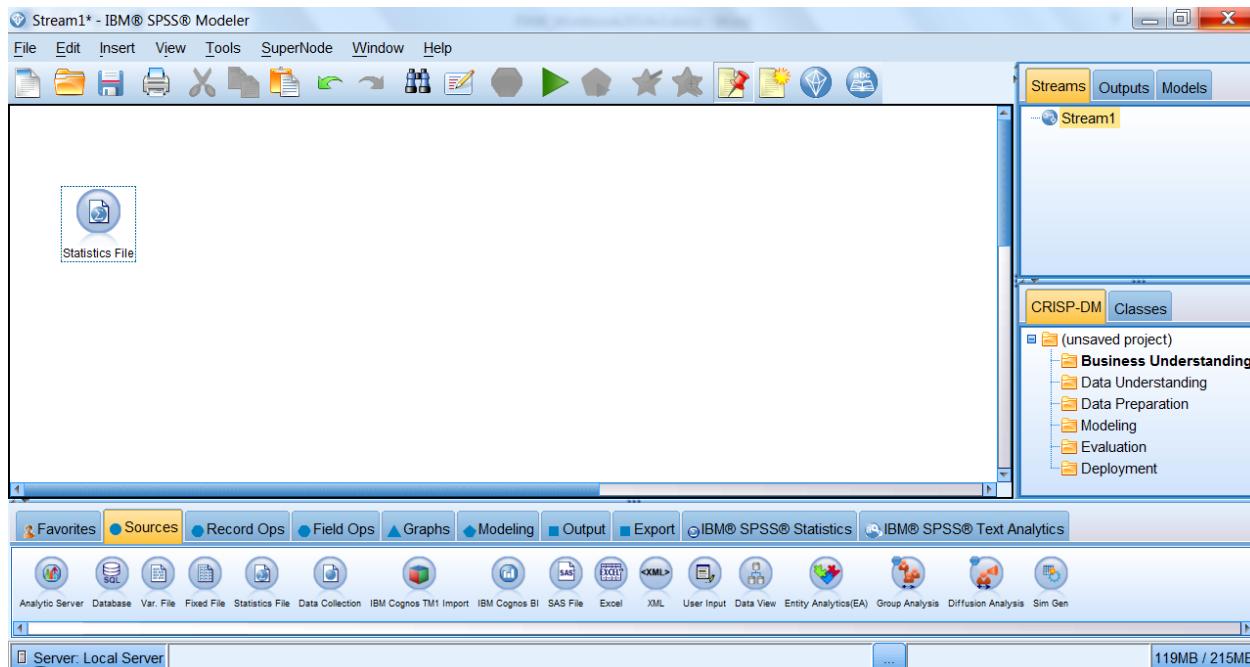
25% lower costs by replacing mass marketing with targeted campaigns

99% faster reports, with the ability to analyze data in minutes instead of weeks

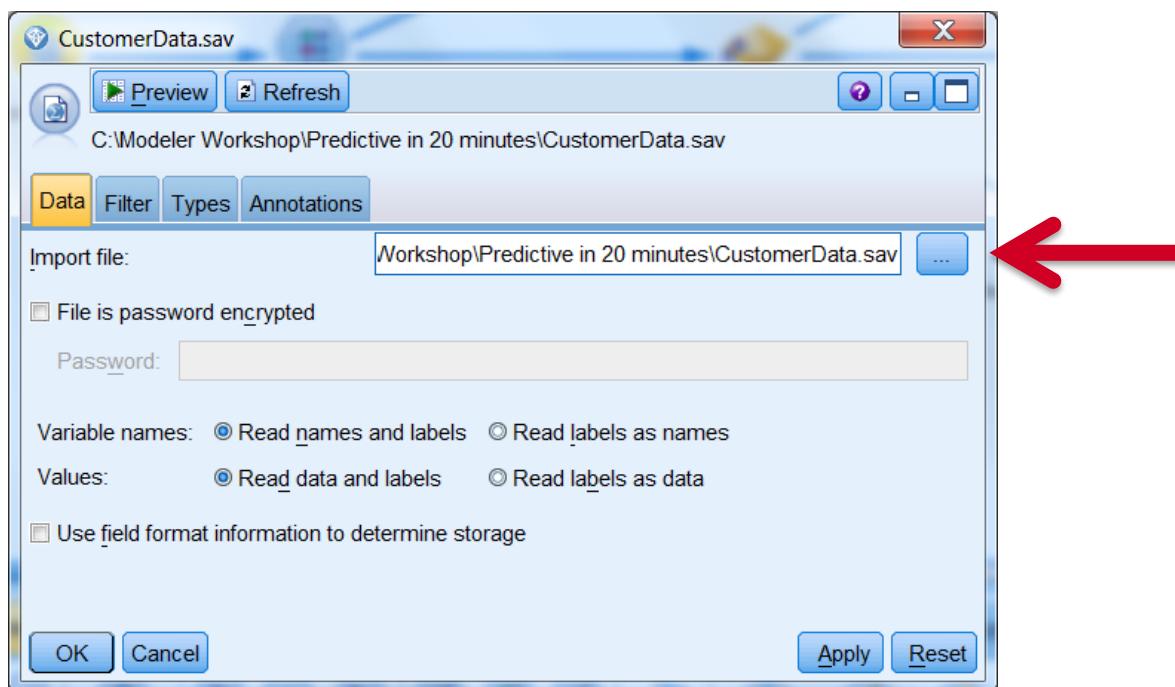
“The ability to analyze and predict customer behavior has become a necessity for retailers who rely on customer loyalty to stay ahead.”

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.



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.

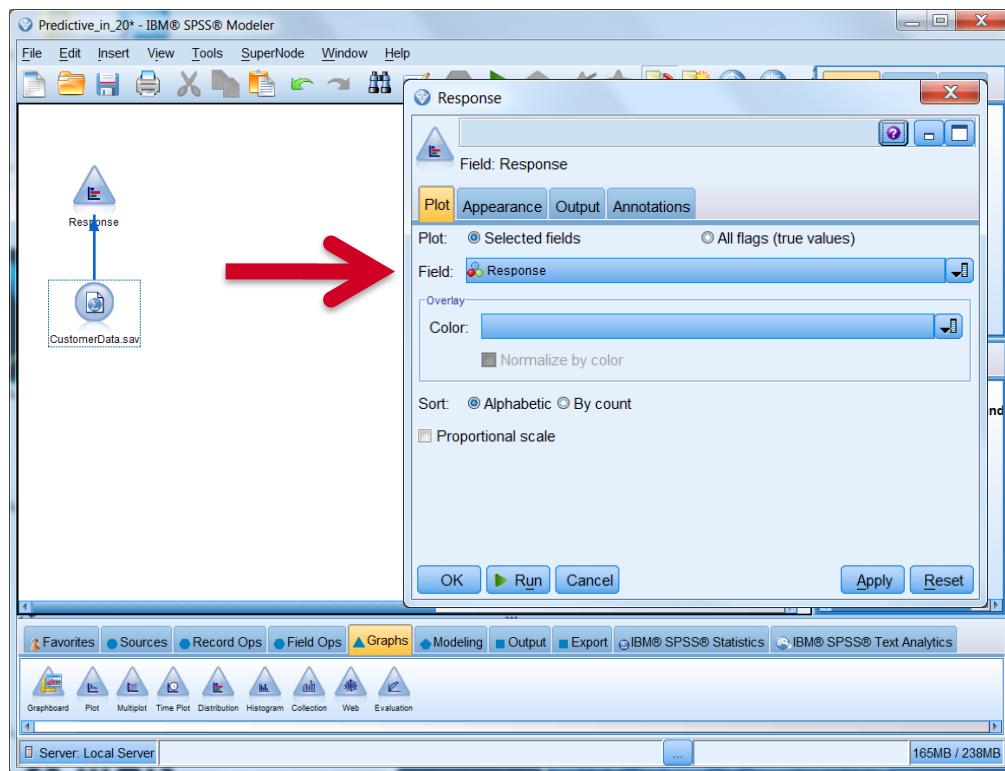


4. Click on the Preview button at the top of the dialog box to see the first 10 records in the file, an extraction of data from a retail company's CRM system. It includes historical data related to their customers' demographics, purchasing behavior, segment, and marketing campaign response.

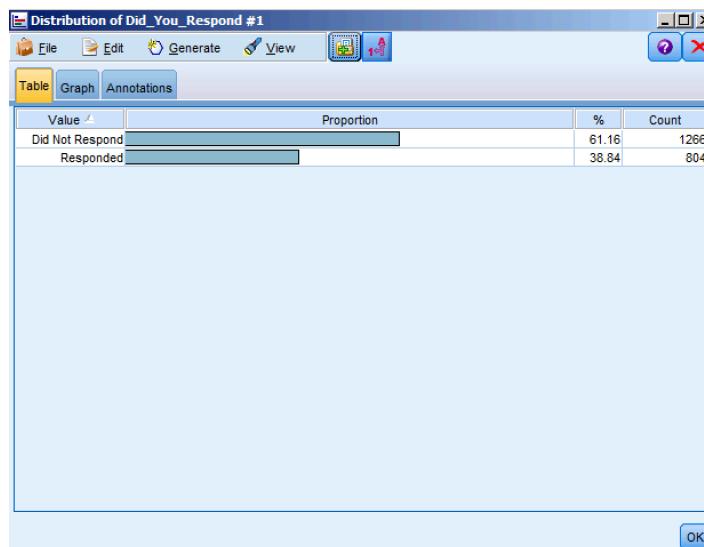
| ID | Sex | Age | Region | Children | Est_Income | Car_Owner | Status | Paymethod | LocalBilltype | Customer_Segments | Trips_To_The_Website_Yr | Avg_Spen |
|----|----------|--------|--------|----------|-------------|-----------|--------|-----------|-----------------------|-------------------|-------------------------|----------|
| 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.

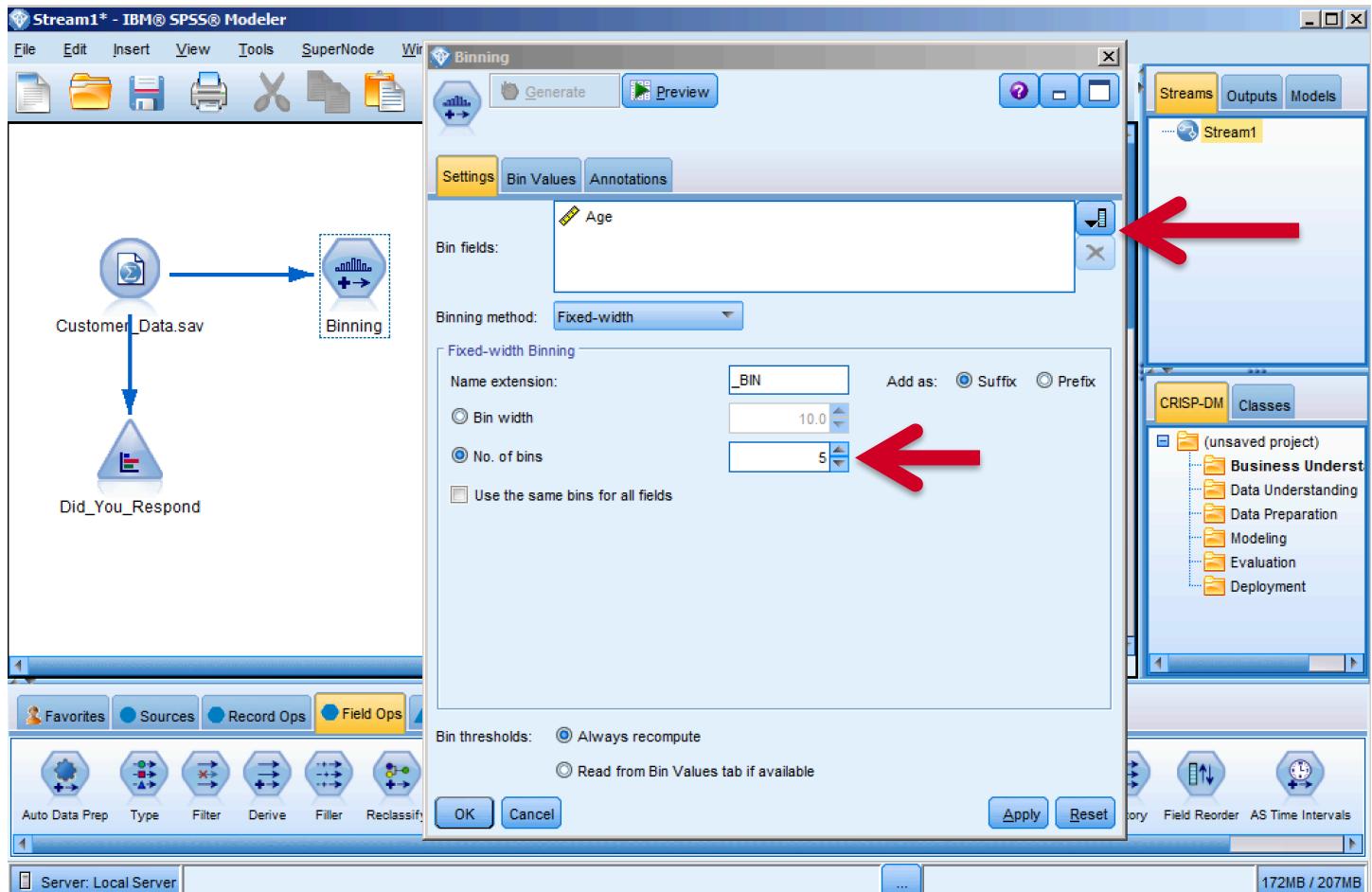


- The resulting graph shows that of the 2070 customers in this dataset, only 38.84%, or 804 customers, responded to the campaign. The remaining 61.16%, or 1266 customers, did not respond. Our goal, then, is to build a model to understand the relationships within the data that lent themselves to customer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



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.



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.

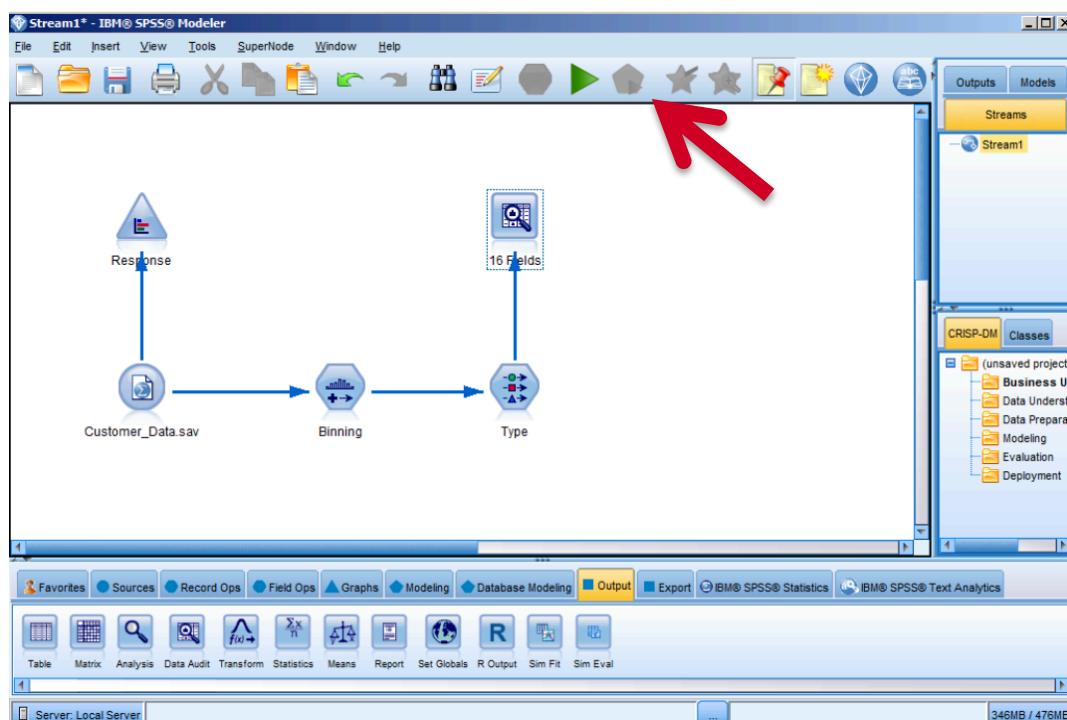
| Field | Measurement | Values | Missing | Check | Role |
|--------------------|-------------|-------------------------|-------------|-------------|---------------|
| ID | Continuous | [1.0,2070.0] | | None | Record ID |
| Sex | Nominal | F,M | | None | Input |
| Age | Continuous | [12.326667,77...] | | 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 | None | Target |
| Age_BIN | Nominal | 1,2,3,4,5 | | None | Input |

View current fields View unused field settings

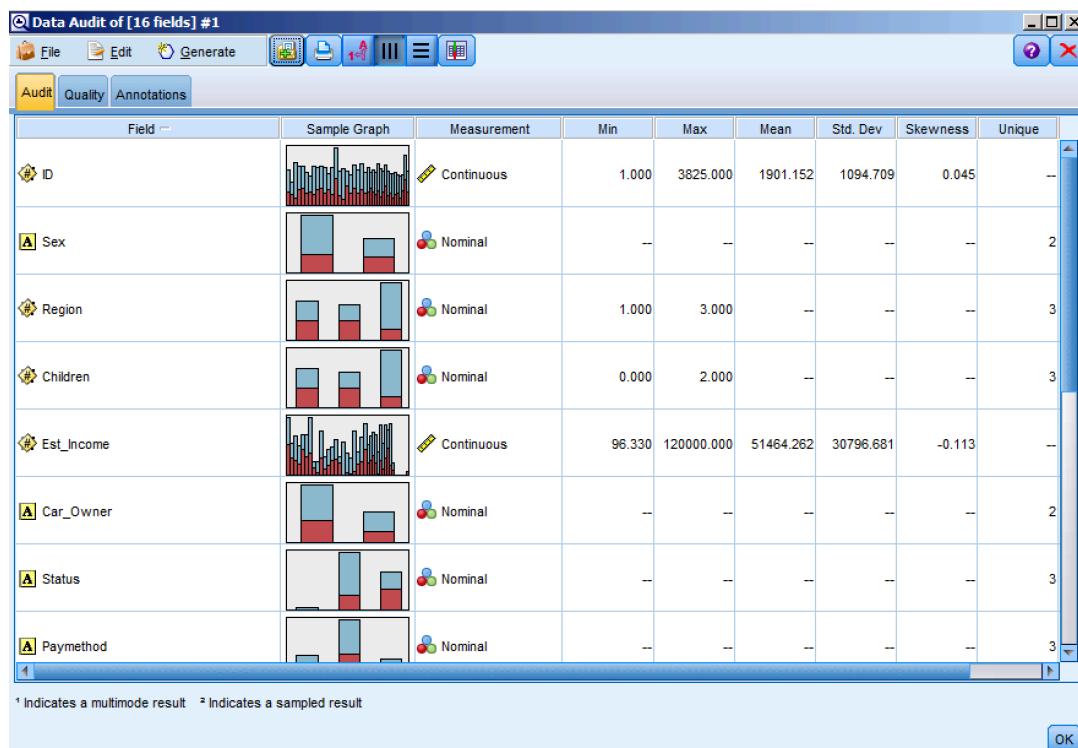
OK Cancel Apply Reset

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.

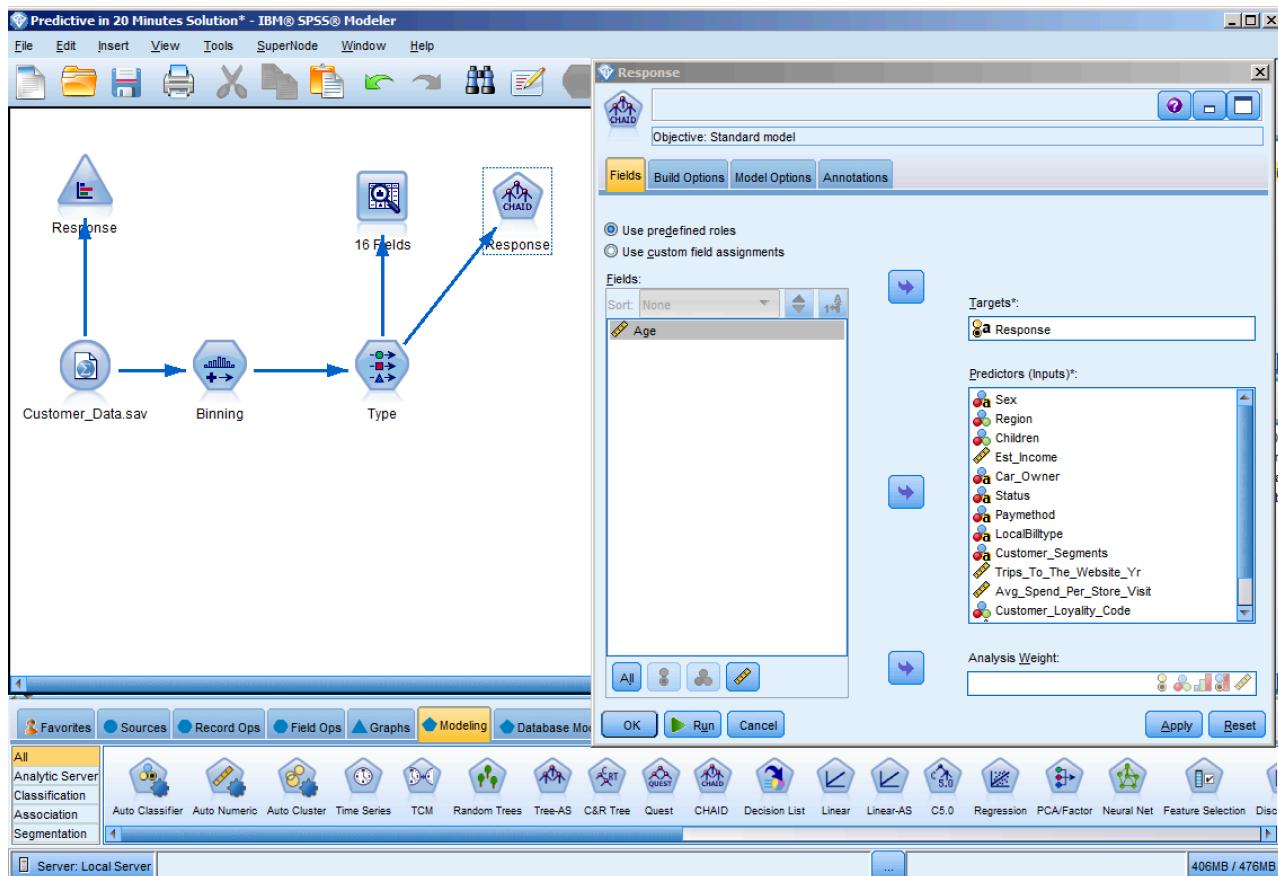


11. In the Audit tab of the resulting output, thumbnail graphs, storage icons, and summary statistics for all fields can be found. Double-clicking on any of the graphs will provide a more detailed outlook of the Field. In the Quality tab (not shown), information about outliers, extremes and missing values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



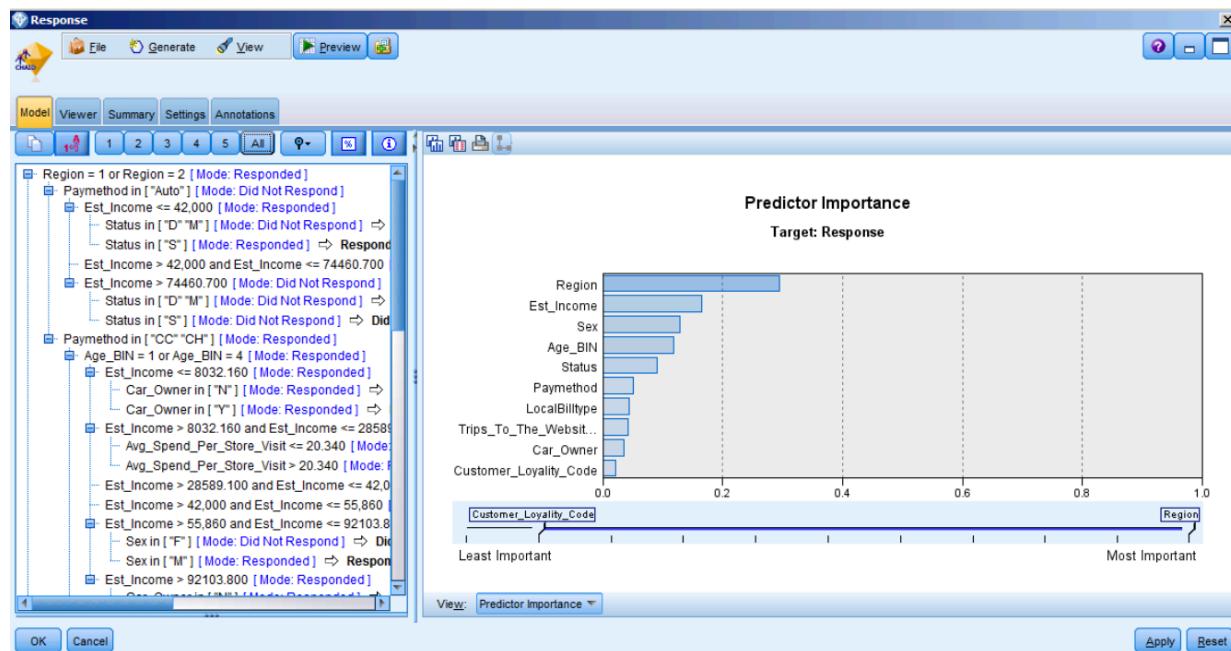
12. Now that we have explored our data, we can build a model to uncover the key drivers resulting in campaign response. We will do this by connecting a CHAID node from the Modeling palette to the Type node.

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.

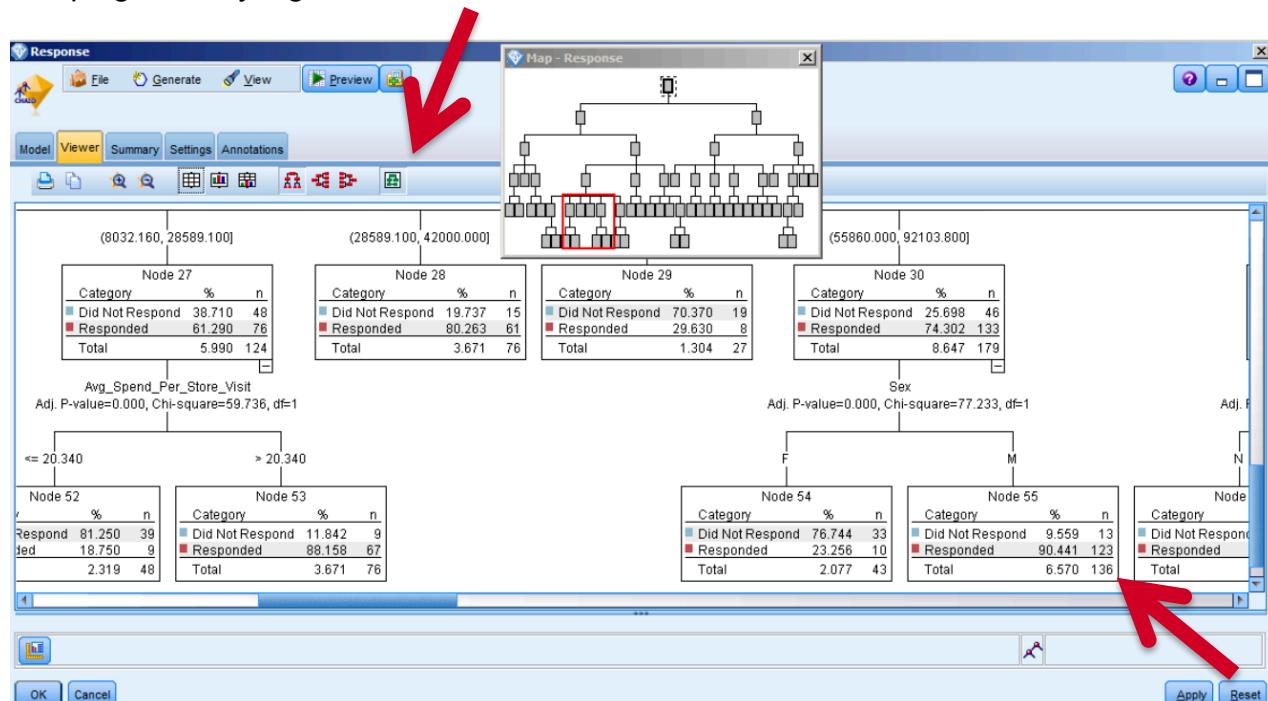


13. The CHAID model is automatically generated and added to our canvas (not shown).

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.

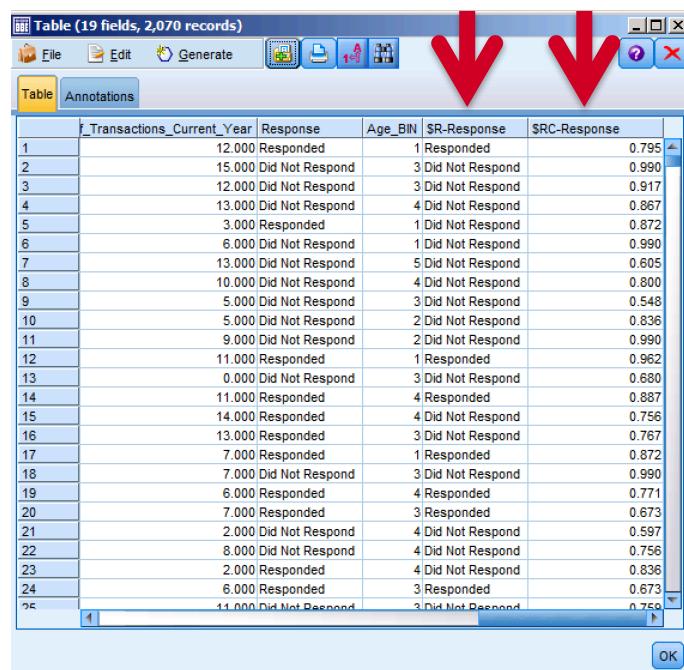


14. The Viewer tab displays the resulting tree, reflecting cut points of the important predictors as determined by the model. To navigate the tree map, click on the “show or hide tree map” icon in the toolbar. Following one branch down, we are able to discover key insights. For example, males who earned 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.



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.

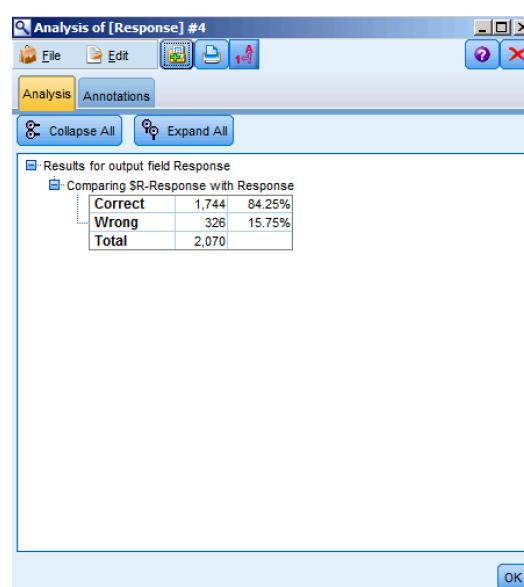


The screenshot shows a 'Table (19 fields, 2,070 records)' window. The table has columns: f_Transactions_Current_Year, Response, Age_BIN, \$R-Response, and SRC-Response. Red arrows point to the \$R-Response and SRC-Response columns. The data includes various transaction counts, response types (Responded, Did Not Respond), age bins, and confidence scores ranging from 0.605 to 0.990.

| | f_Transactions_Current_Year | Response | Age_BIN | \$R-Response | SRC-Response |
|----|-----------------------------|-----------------|---------|-------------------|--------------|
| 1 | 12.000 | Responded | 1 | 1 Responded | 0.795 |
| 2 | 15.000 | Did Not Respond | 3 | 3 Did Not Respond | 0.990 |
| 3 | 12.000 | Did Not Respond | 3 | 3 Did Not Respond | 0.917 |
| 4 | 13.000 | Did Not Respond | 4 | 4 Did Not Respond | 0.867 |
| 5 | 3.000 | Responded | 1 | 1 Did Not Respond | 0.872 |
| 6 | 6.000 | Did Not Respond | 1 | 1 Did Not Respond | 0.990 |
| 7 | 13.000 | Did Not Respond | 5 | 5 Did Not Respond | 0.605 |
| 8 | 10.000 | Did Not Respond | 4 | 4 Did Not Respond | 0.800 |
| 9 | 5.000 | Did Not Respond | 3 | 3 Did Not Respond | 0.548 |
| 10 | 5.000 | Did Not Respond | 2 | 2 Did Not Respond | 0.836 |
| 11 | 9.000 | Did Not Respond | 2 | 2 Did Not Respond | 0.990 |
| 12 | 11.000 | Responded | 1 | 1 Responded | 0.962 |
| 13 | 0.000 | Did Not Respond | 3 | 3 Did Not Respond | 0.680 |
| 14 | 11.000 | Responded | 4 | 4 Responded | 0.887 |
| 15 | 14.000 | Responded | 4 | 4 Did Not Respond | 0.756 |
| 16 | 13.000 | Responded | 3 | 3 Did Not Respond | 0.767 |
| 17 | 7.000 | Responded | 1 | 1 Responded | 0.872 |
| 18 | 7.000 | Did Not Respond | 3 | 3 Did Not Respond | 0.990 |
| 19 | 6.000 | Responded | 4 | 4 Responded | 0.771 |
| 20 | 7.000 | Responded | 3 | 3 Responded | 0.673 |
| 21 | 2.000 | Did Not Respond | 4 | 4 Did Not Respond | 0.597 |
| 22 | 8.000 | Did Not Respond | 4 | 4 Did Not Respond | 0.756 |
| 23 | 2.000 | Responded | 4 | 4 Did Not Respond | 0.836 |
| 24 | 6.000 | Responded | 3 | 3 Responded | 0.673 |
| 25 | 11.000 | Did Not Respond | 3 | 3 Did Not Respond | 0.750 |

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.

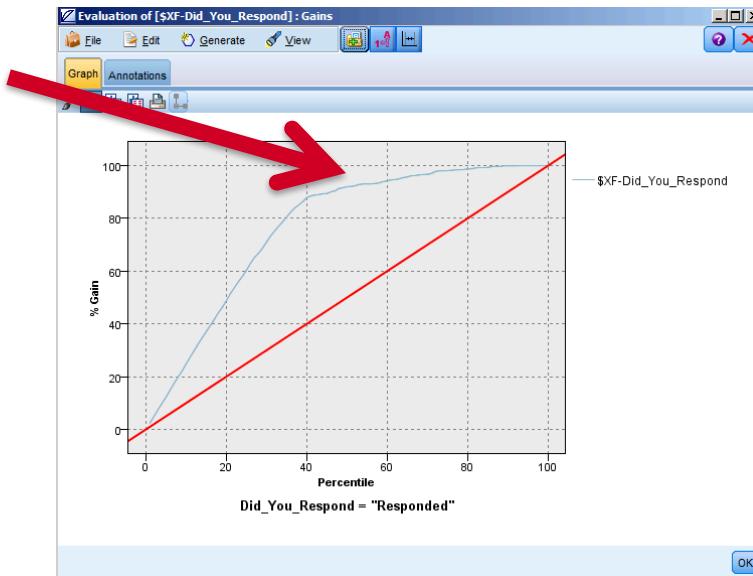


The screenshot shows an 'Analysis of [Response] #4' window. It displays a table comparing \$R-Response with Response. The table has columns: Correct, Wrong, and Total. The data shows 1,744 correct predictions (84.25%), 326 wrong predictions (15.75%), and a total of 2,070 records.

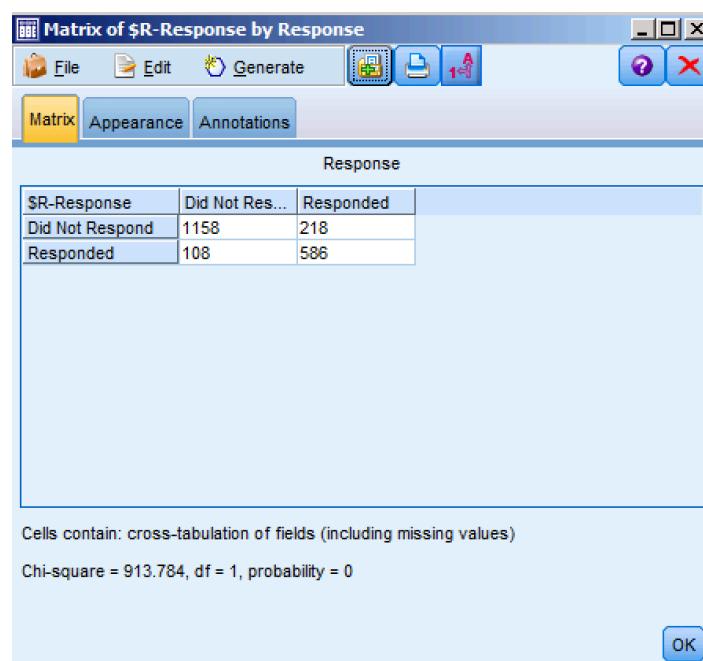
| | | |
|---------|-------|--------|
| Correct | 1,744 | 84.25% |
| Wrong | 326 | 15.75% |
| Total | 2,070 | |

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.



- 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.



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.

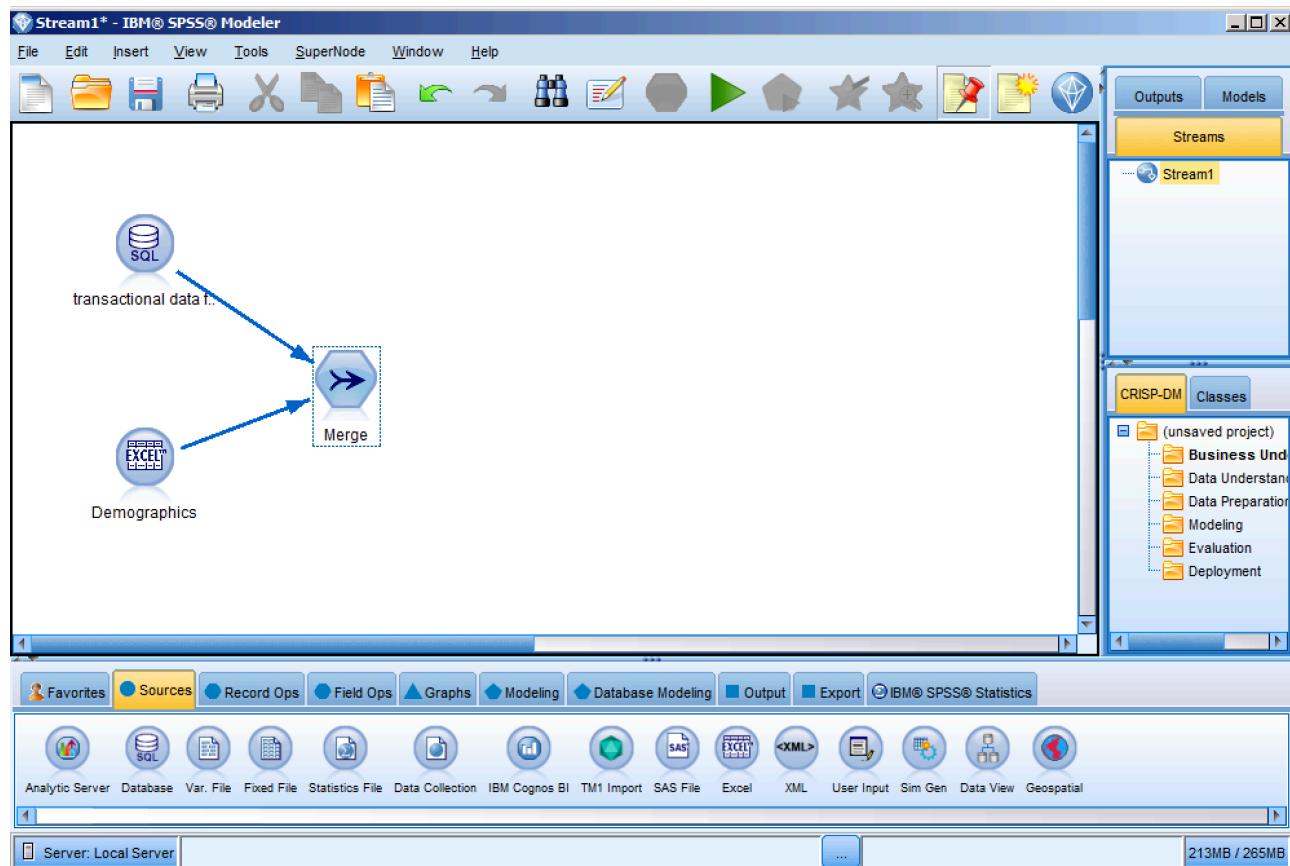
Predicts success of a new show six weeks in advance of its release and adjusts marketing accordingly.

"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.

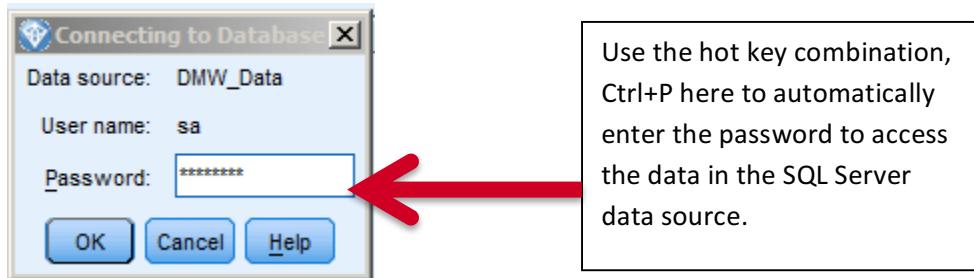


You might recall that in the Predictive in 20 Minutes exercise there was a field assigning each customer to a marketing cluster or segment. We're going to take a step back in our story to take a deeper look at how that is accomplished in IBM SPSS Modeler.

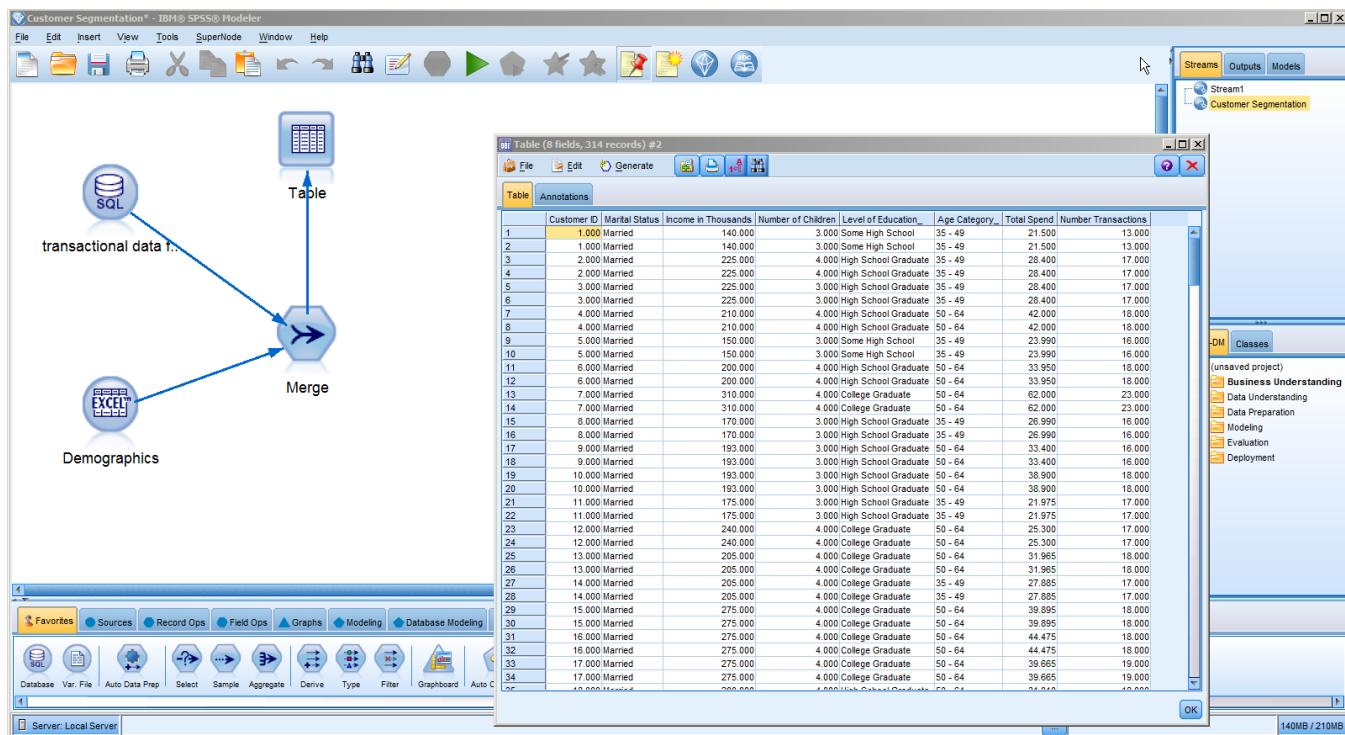
We will be joining two data files together; one file contains customer transaction data, and exists in a database, while the other file is an Excel file and contains customer demographics. Though the files are in different formats, the Merge node in IBM SPSS Modeler can very easily join the files without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

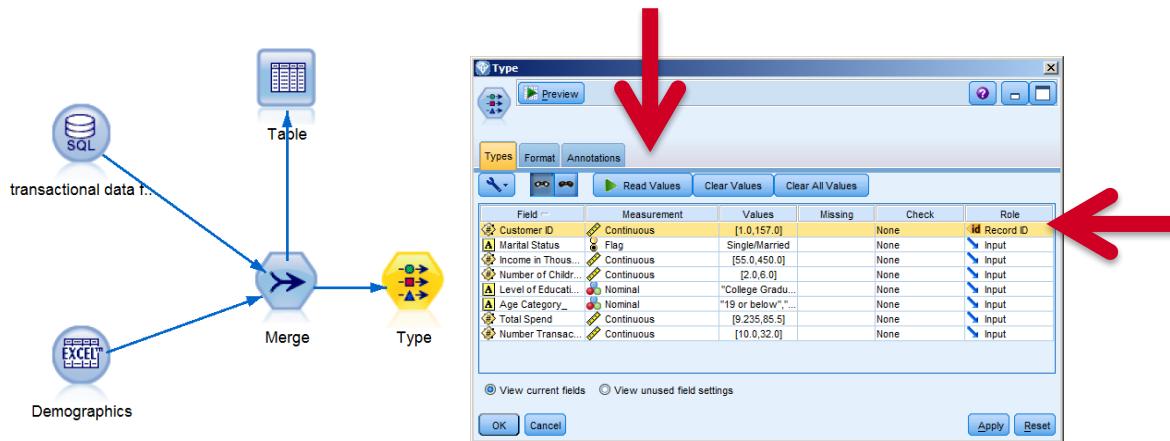
2. From the Output palette, add a Table node to the canvas and connect it to the Merge node.
3. Once connected, right click on the Table node and select Run to review the data after having been merged. If prompted for a password, use “Ctrl+P” to automatically populate the password.



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.

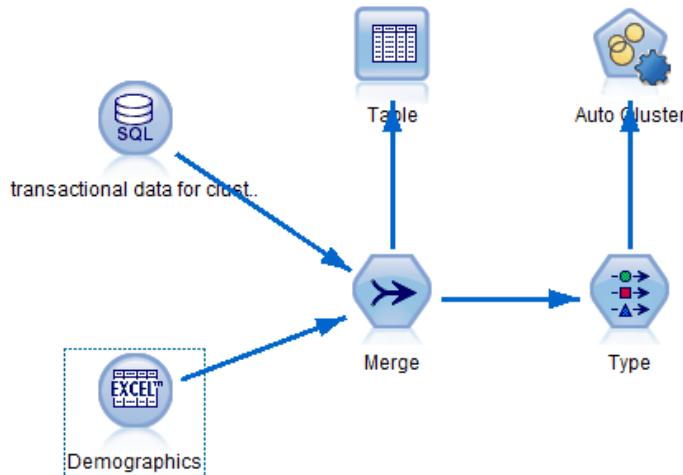


4. From the Field Ops palette at the bottom of the screen, select a Type node and attach it to the Merge node. Double-click on the Type node to edit it. Click on the Read Values push button, and change the role of “Customer ID” to “Record ID” (shown below). Once done, click on the OK button.

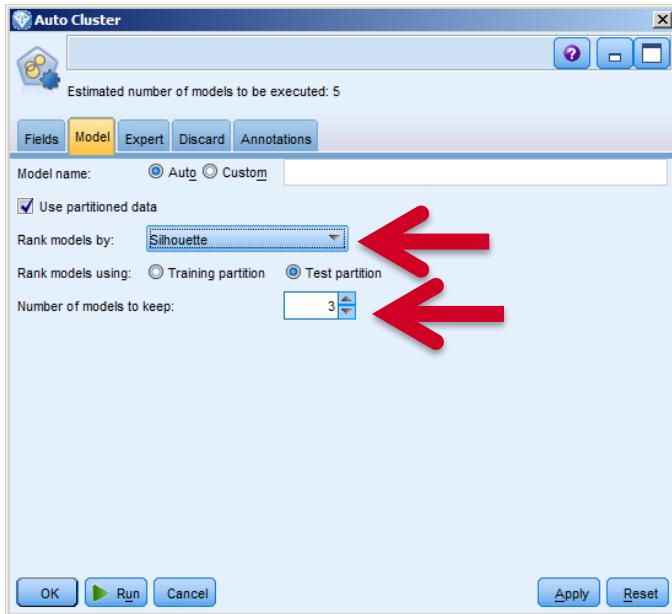


At this point we are ready to cluster the cases into segments. For this we will use the ‘Auto Cluster’ node. The Auto Cluster node allows you to try all of the clustering algorithms and, at your discretion, any or all of their parameters. It builds all of the models you specify and shows you the best 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.



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

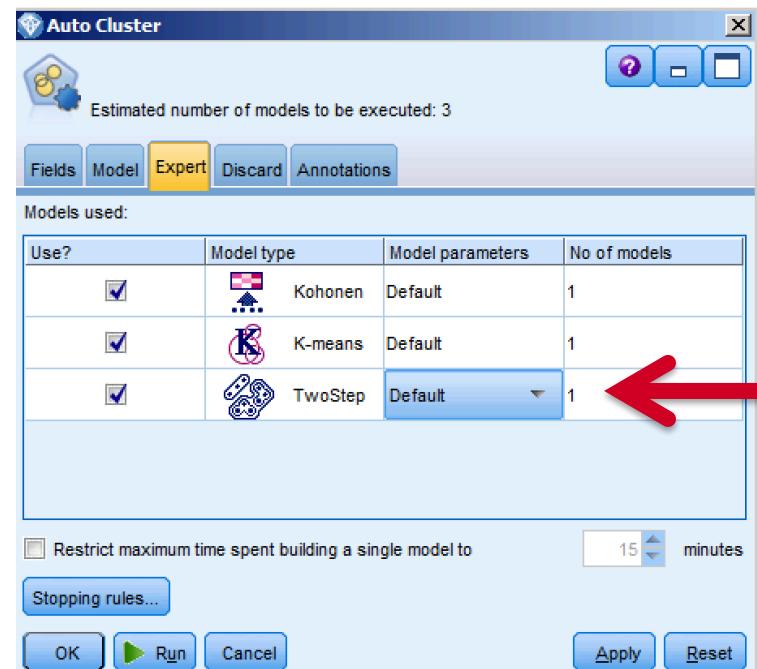
Also keep the default number of models to keep, 3. This means that the 3 best models, based on their silhouette measure, will be retained for our use.

Click on the 'Expert' tab.

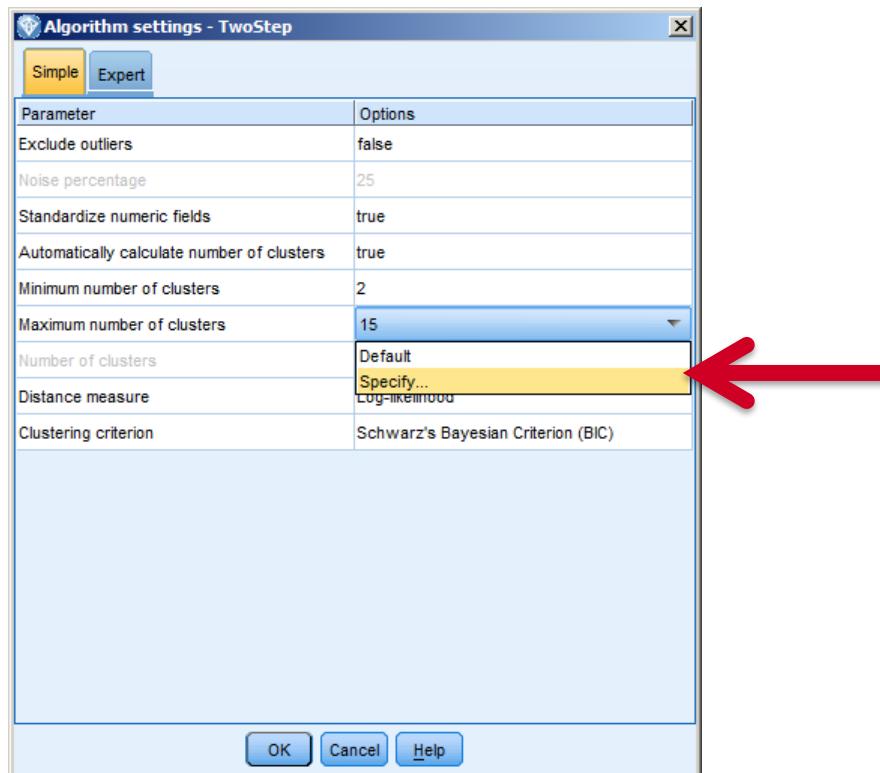
You notice that there are 3 clustering algorithms in the list. Model parameters for each can be specified.

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 will appear as shown in the next section.

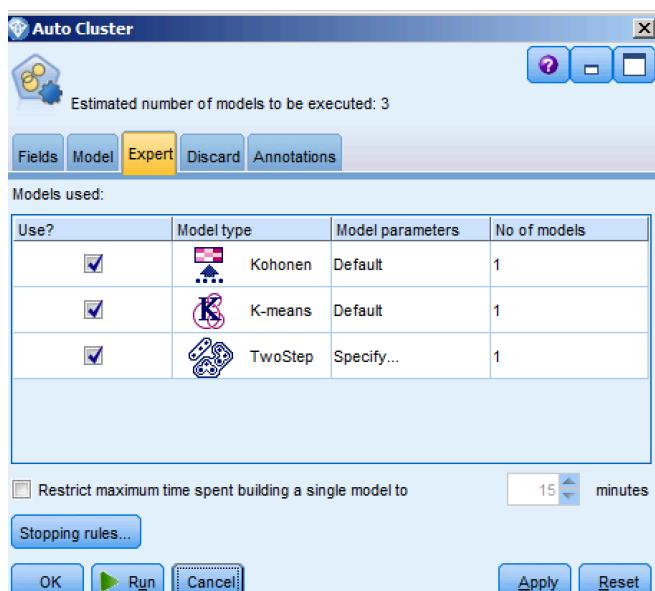


The ‘Maximum number of clusters’ is 15. Click on the number 15, and click **Specify**.



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.



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.

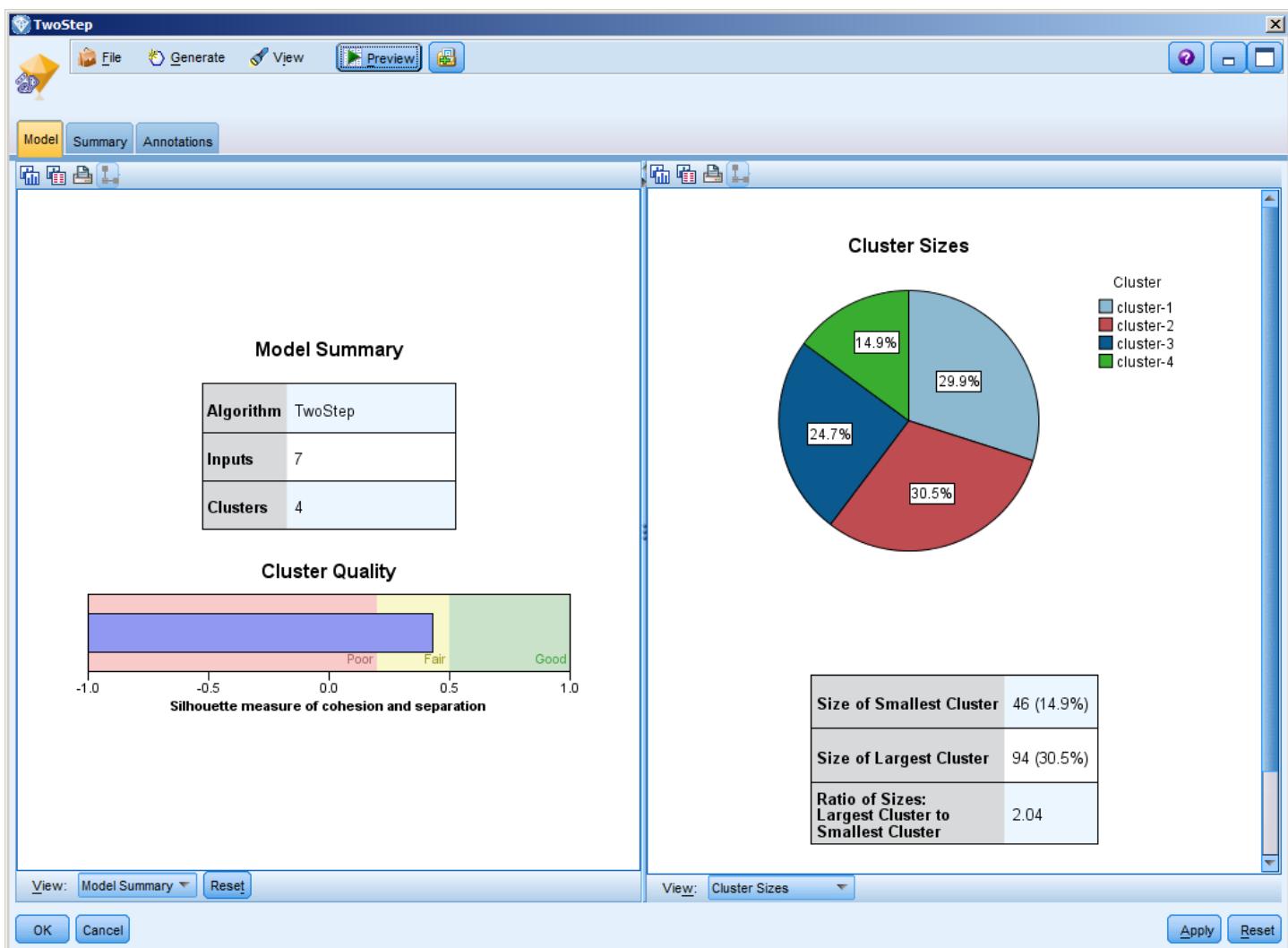
| Use? | Graph | Model | Build Time (mins) | Silhouette | Number of Clusters | Smallest Cluster (N) | Smallest Cluster (%) | Largest Cluster (N) | Largest Cluster (%) | Smallest/Largest | Importance |
|-------------------------------------|---------|-------|-------------------|------------|--------------------|----------------------|----------------------|---------------------|---------------------|------------------|------------|
| <input checked="" type="checkbox"/> | TwoS... | < 1 | | 0.434 | 4 | 46 | 14 | 94 | 30 | 0.489 | 0.0 |
| <input type="checkbox"/> | K-me... | < 1 | | 0.377 | 5 | 2 | 0 | 136 | 43 | 0.015 | 0.0 |
| <input type="checkbox"/> | 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.

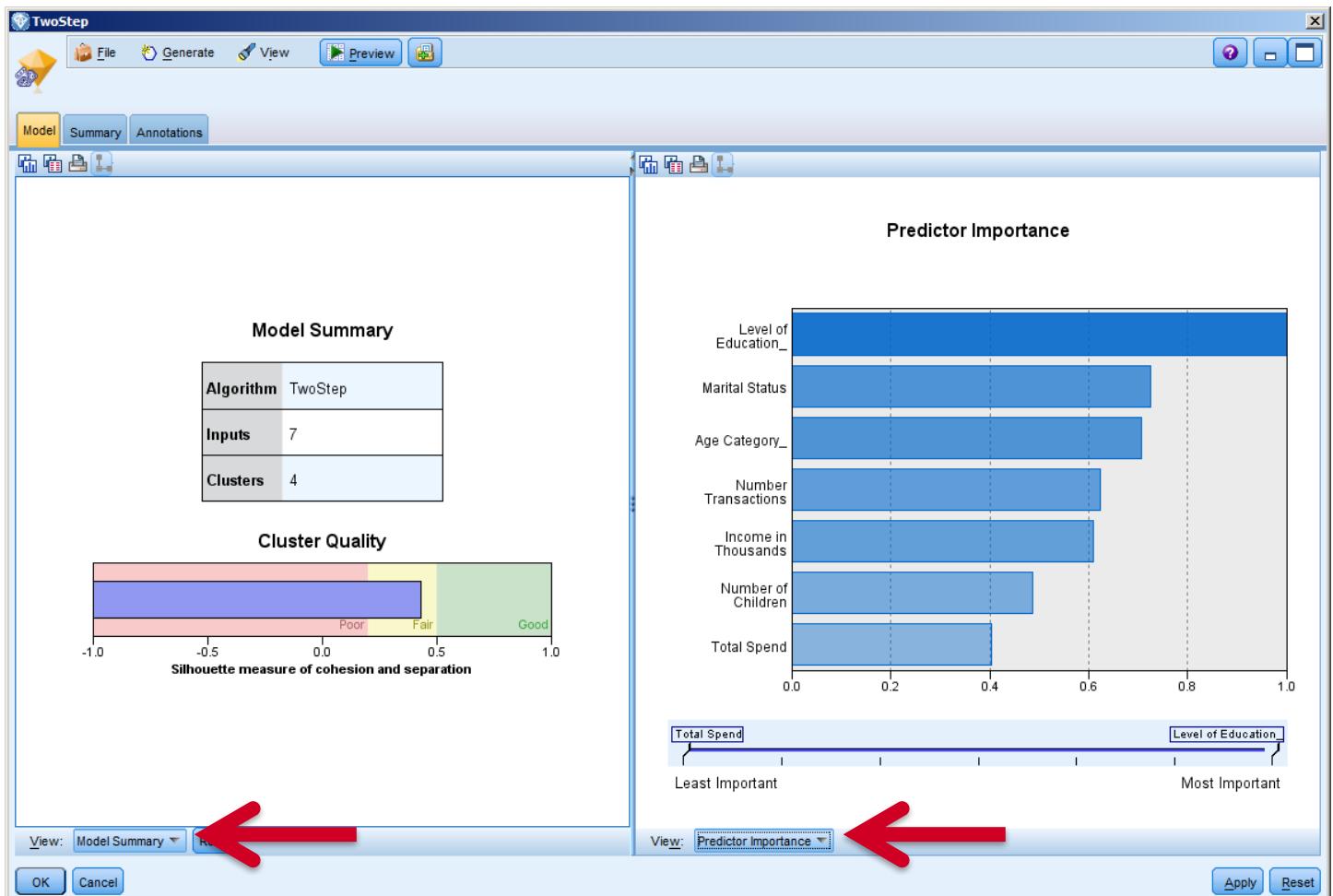


We specified a range of 2 to 4 clusters; and the Two Step clustering engine resolved into 4 clusters.

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.

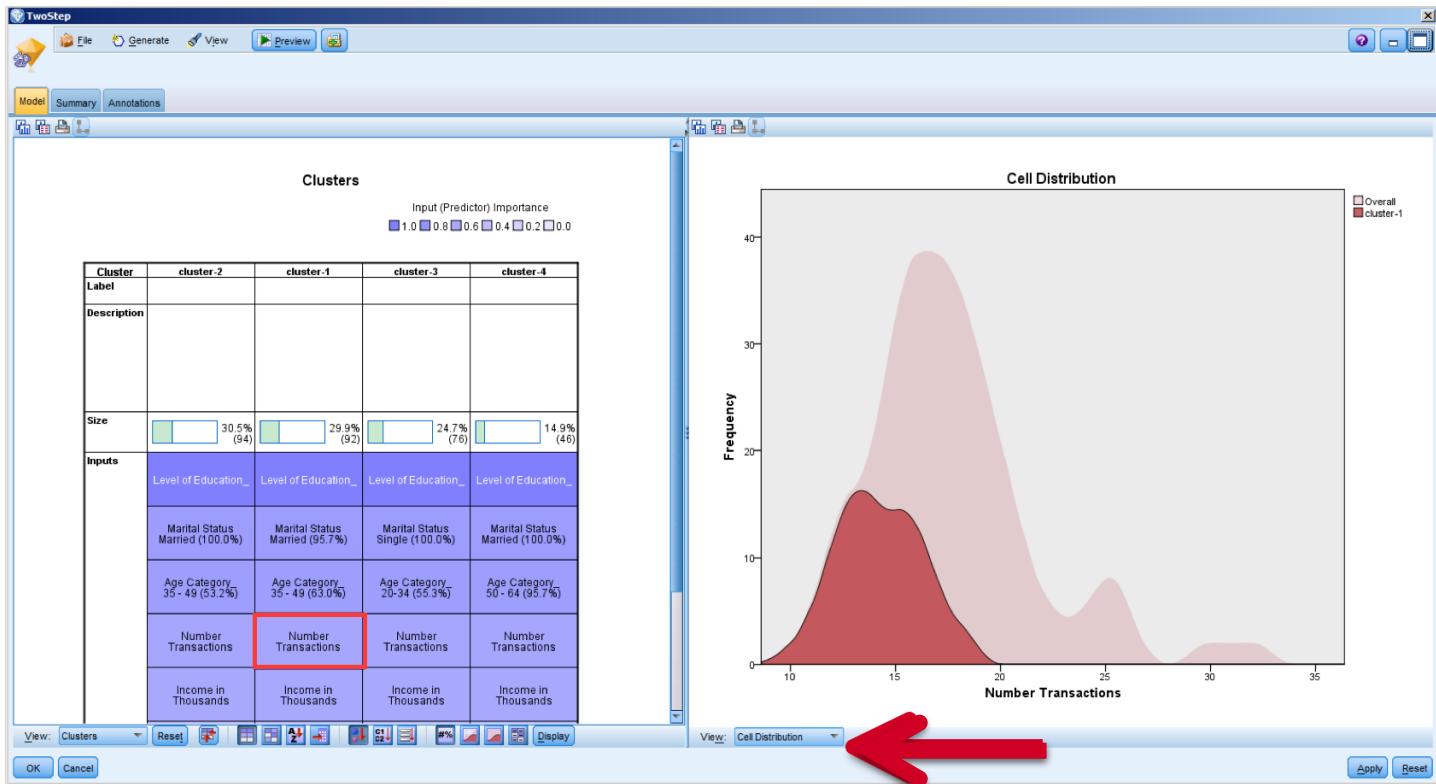
On the right side of the viewer is a pie chart illustrating the cluster sizes.

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



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.

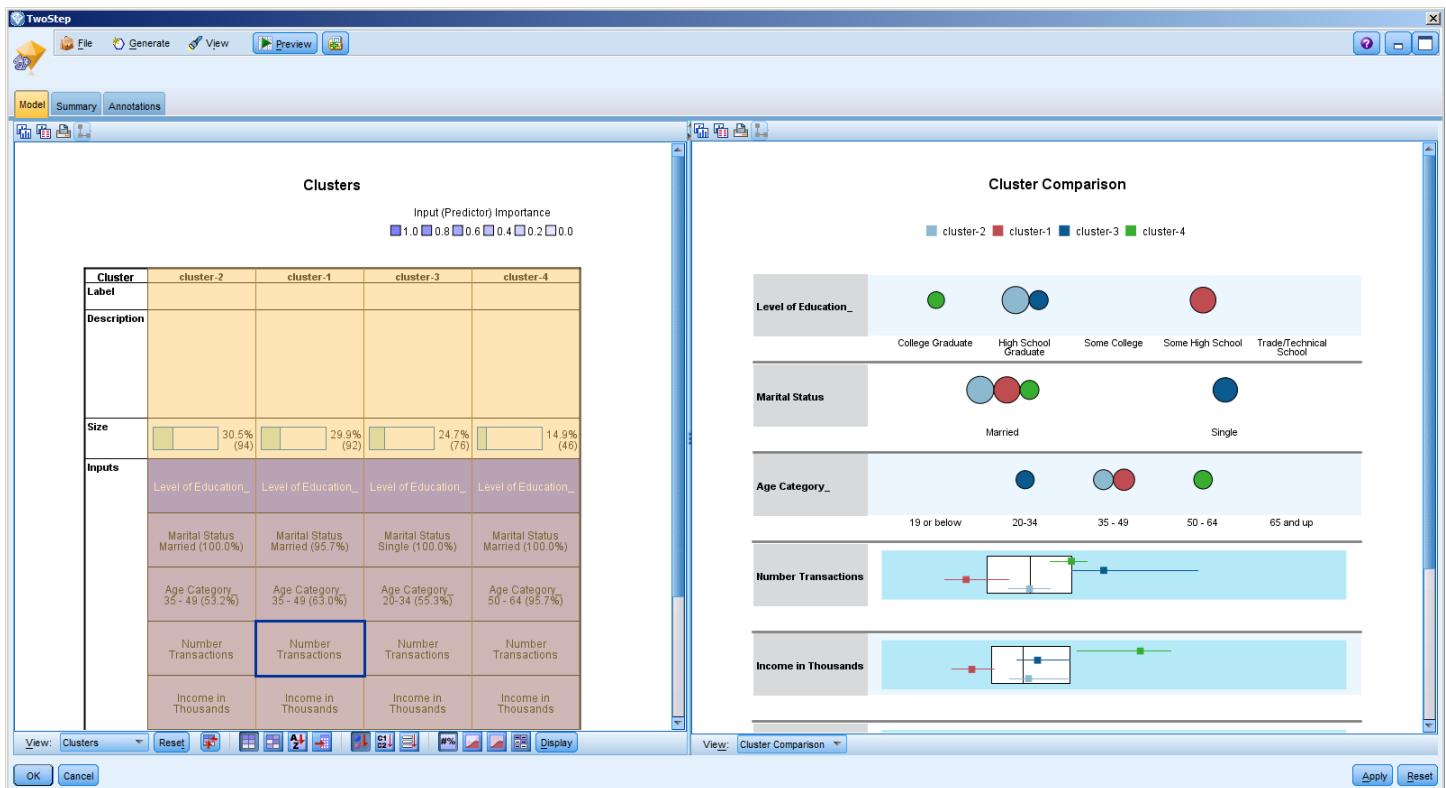
9. Now, from the drop-down menu in the left viewer, select Clusters; from the drop-down menu in the right viewer, select Cell Distribution (shown below in red).



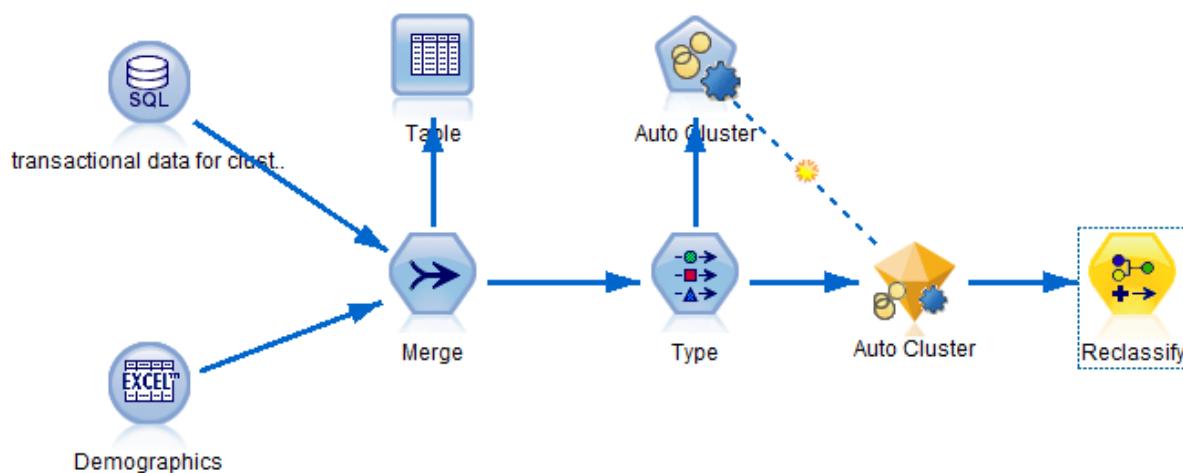
The left panel of the Viewer displays the clusters in order of their size, left to right. The darkness of the shading of each variable indicates its importance in cluster definition; the lighter the shading, the less important is the variable in defining the clusters.

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.



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).

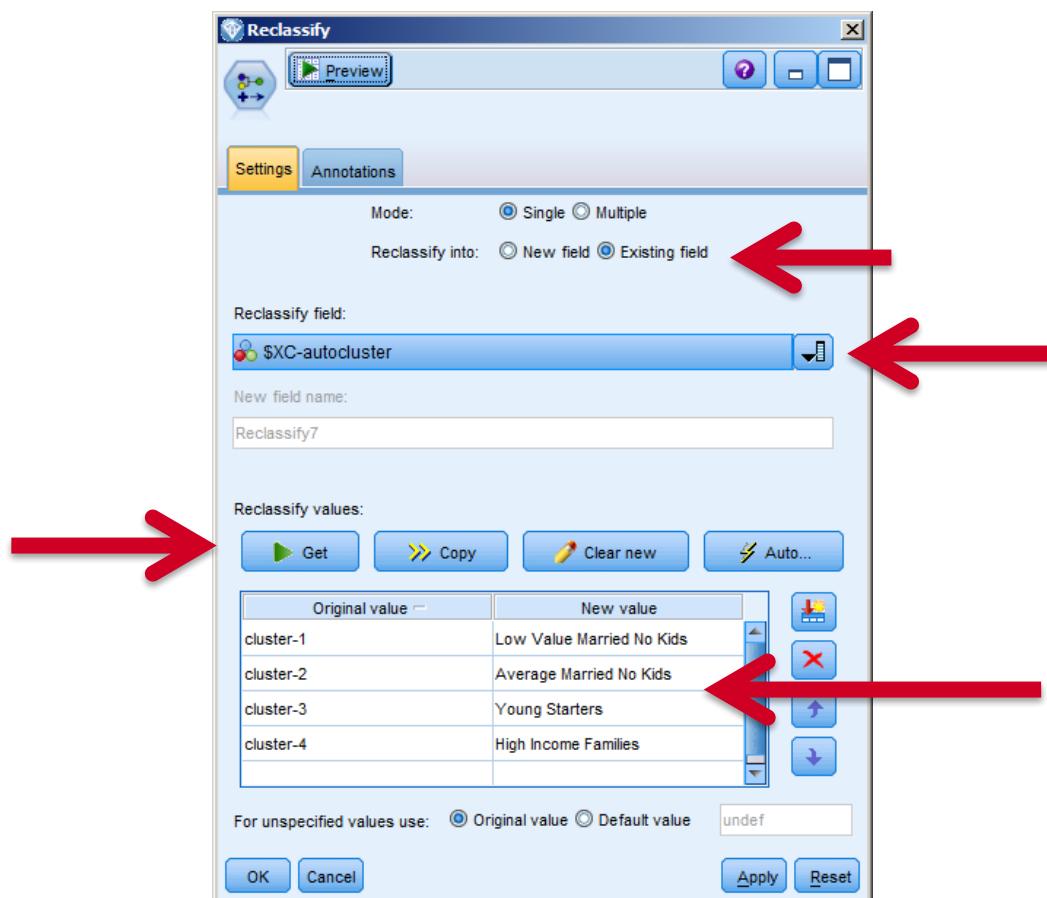


14. Double-click on the Reclassify node to edit it.

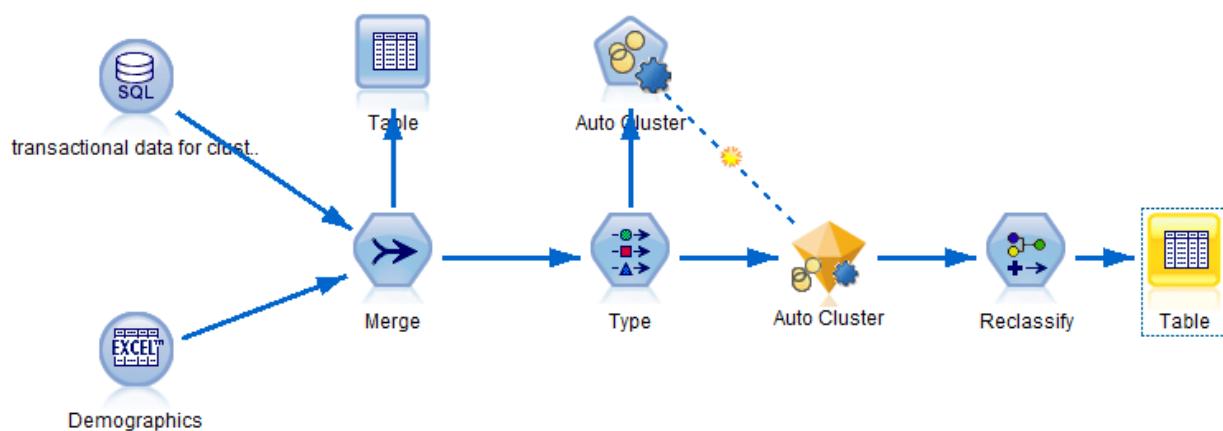
In the Settings tab of the Reclassify dialog, use the drop-down menu to select the variable **\$XC-autocluster** as the field to be reclassified.

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.



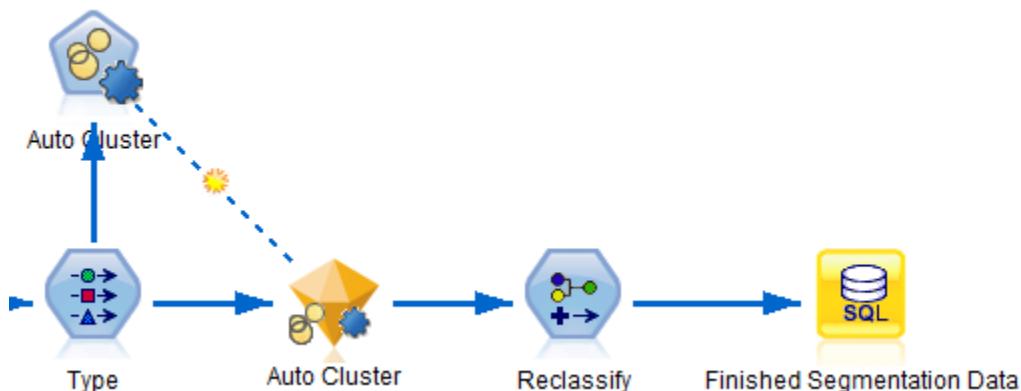
15. From the Output palette, drag a Table node onto the workbench and connect it to the Reclassify node. Once connected, right-click on the Table node and Select Run.



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

| Table (9 fields, 314 records) | | | | | | | |
|-------------------------------|---------------------|--------------------|----------------------|---------------|-------------|---------------------|---------------------------|
| | 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.

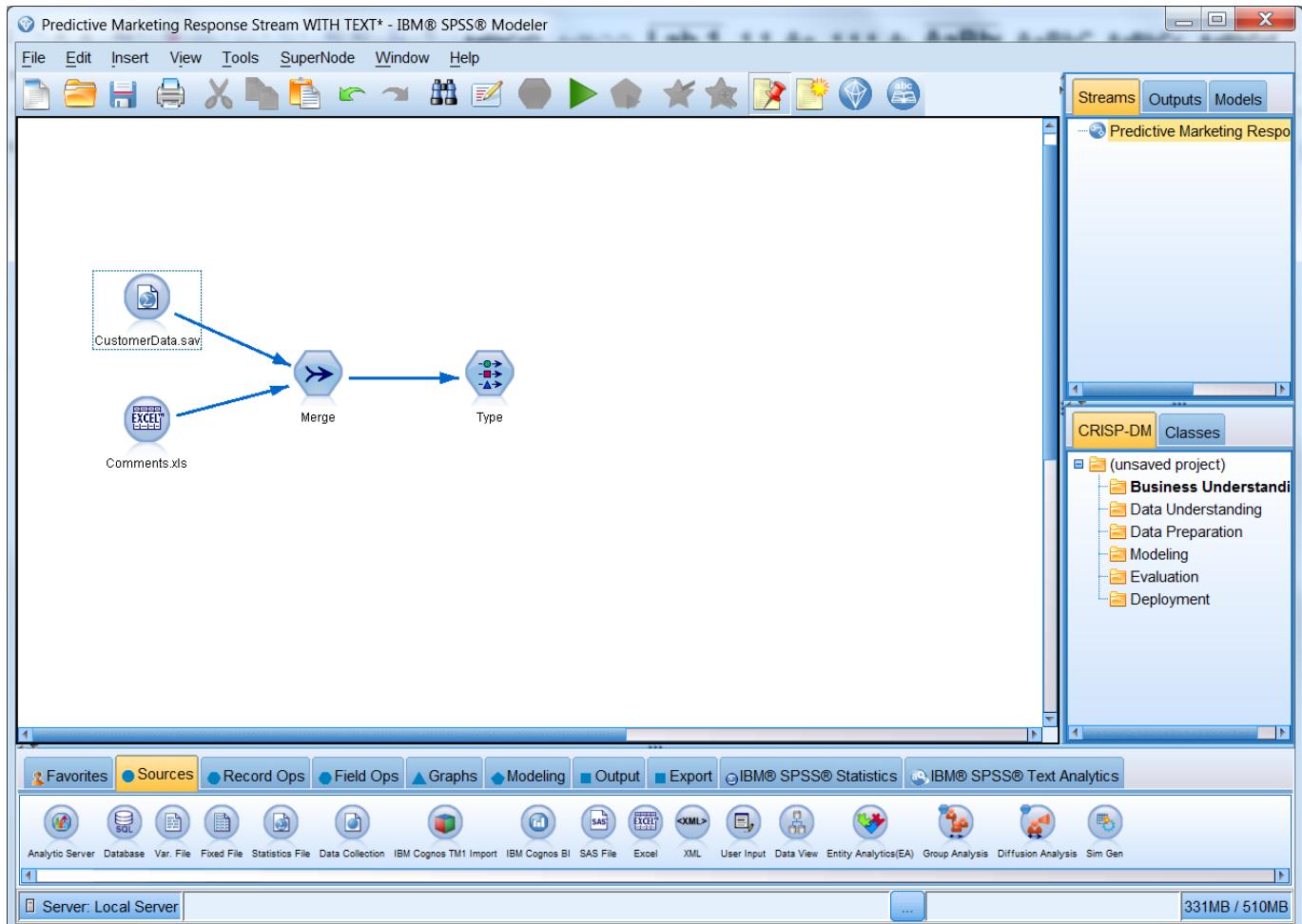
Enabled the company to evaluate more than 450 variables for predicting customer defection within 90 days.

“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 “Classification 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.

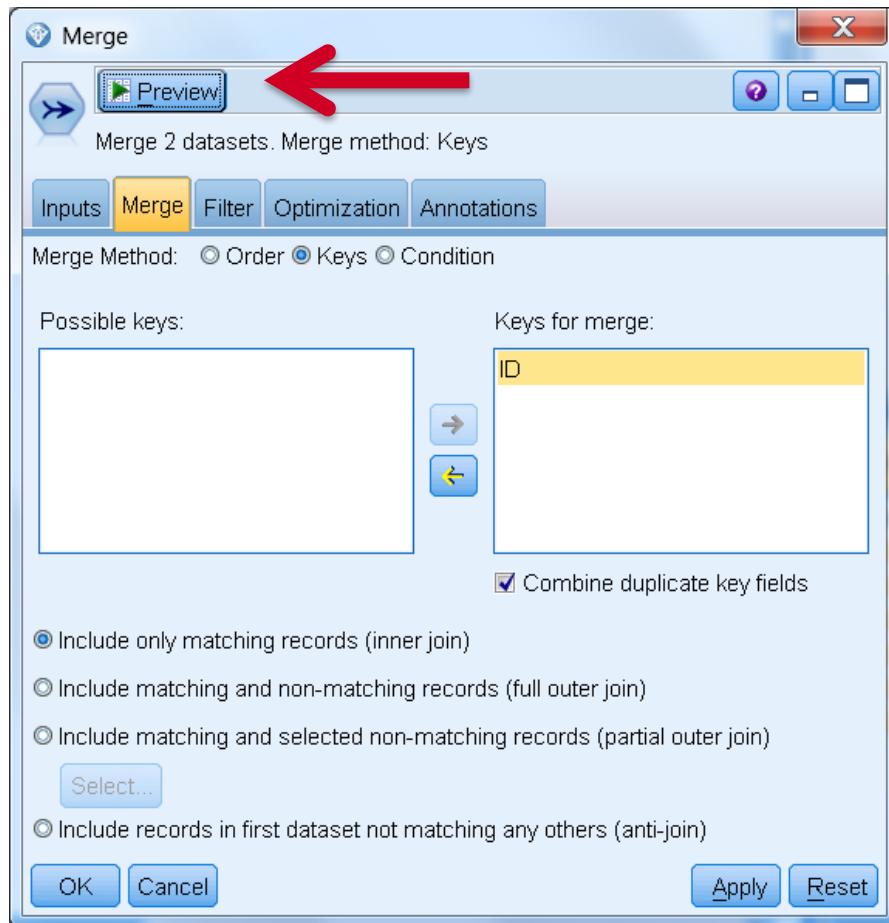


This exercise begins with a stream that has already been partially constructed. Note, in particular, that we are using the same customer data from the Predictive in 20 Minutes exercise. Notice also that we will be using unstructured data in the form of customer comments, and contained in a Microsoft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



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)

| | Loyalty_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.

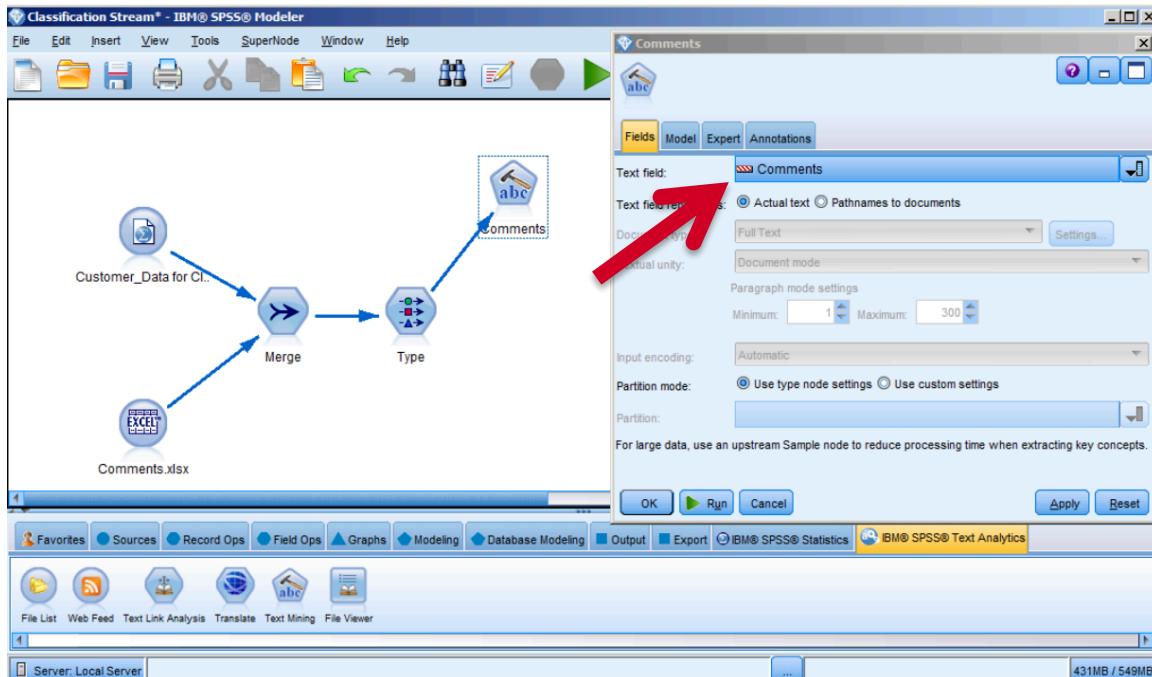
| Field | Measurement | Values | Missing | Check | Role |
|-------------------|-------------|------------------|---------|-------|-----------|
| # ID | Continuous | [1.0,1620.0] | | None | Record ID |
| A Sex | Nominal | F,M | | None | Input |
| # Region | Nominal | 1.0,2.0,3.0 | | None | Input |
| # Children | Nominal | 0.0,1.0,2.0 | | None | Input |
| # Est_Income | Continuous | [96.33,1200...] | | None | Input |
| A Car_Owner | Nominal | N,Y | | None | Input |
| A Status | Nominal | D,M,S | | None | Input |
| A Paymethod | Nominal | Auto,CC,CH | | None | Input |
| A LocalBilltype | Nominal | Budget,Free... | | None | Input |
| A 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 |
| A Response | Flag | Responded/... | | None | Input |
| # Age_BIN | Nominal | 1.0,2.0,3.0,4... | | None | Input |
| A Comments | Typeless | | | None | None |

View current fields View unused field settings

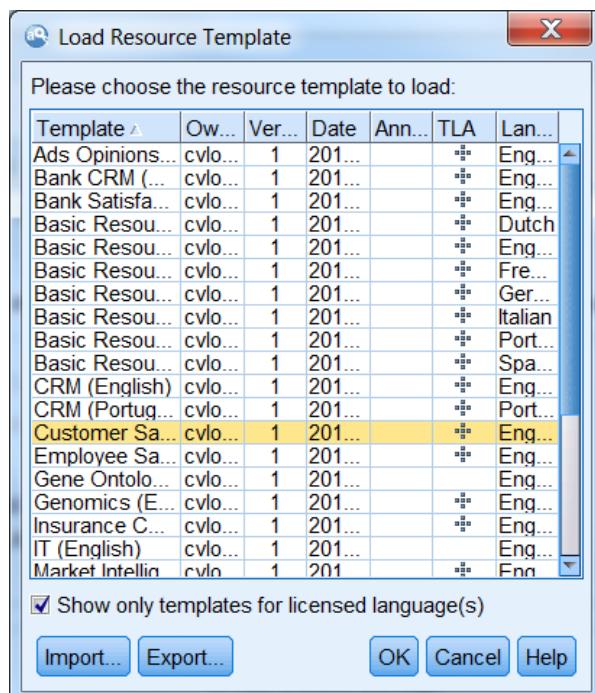
Using Text Analytics, we can identify patterns in the unstructured data and from them create categories which contain the *ideas* and *sentiments* as expressed by customers.

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.



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.



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.

The screenshot shows the IBM SPSS Text Analytics for Modeler Interactive Workbench interface. The top menu bar includes File, Edit, View, Generate, Categories, Tools, and Help. The top right corner has a "Categories and Concepts" dropdown. The interface is divided into several panes:

- Category Bar:** Shows categories like All Documents, Uncategorized, and No concepts extracted. A note says: "To show bars or graphs after building categories, make a selection in another pane and click Display."
- Category Web:** Not visible in the screenshot.
- Category Web Table:** Not visible in the screenshot.
- Extract:** This pane displays 536 concepts in a table format. The columns are Concept, In, Global, Docs, and Type. Some examples from the table include:

| Concept | In | Global | Docs | Type |
|-------------|---------|---------|---------|-------------------|
| like | 61 (3%) | 51 (6%) | 51 (6%) | <Positive> |
| easy to use | 52 (2%) | 50 (6%) | 50 (6%) | <Positive> |
| sound | 50 (2%) | 48 (5%) | 48 (5%) | <Features> |
| nothing | 46 (2%) | 46 (5%) | 46 (5%) | <Uncertain> |
| good | 46 (2%) | 45 (5%) | 45 (5%) | <Positive> |
| portable | 44 (2%) | 43 (5%) | 43 (5%) | <Positive> |
| expensive | 43 (2%) | 43 (5%) | 43 (5%) | <NegativeBudget> |
| excellent | 46 (2%) | 41 (5%) | 41 (5%) | <Positive> |
| size | 37 (2%) | 37 (4%) | 37 (4%) | <Characteristics> |
| songs | 39 (2%) | 36 (4%) | 36 (4%) | <Unknown> |
| product | 33 (1%) | 31 (3%) | 31 (3%) | <Products> |
| listening | 31 (1%) | 30 (3%) | 30 (3%) | <Unknown> |
| cool | 21 (1%) | 21 (2%) | 21 (2%) | <Positive> |
- Comments:** This pane lists 41 comments. Examples include:
 - 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 wherever 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 songs from CD to the device.

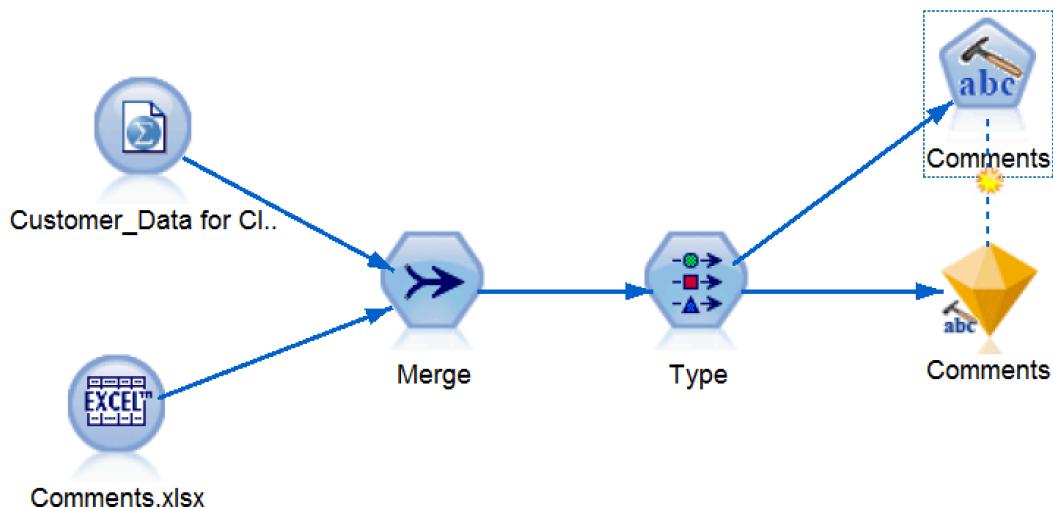
8. To begin the process of building categories, click on Categories > Build Categories > Build now (not shown).

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.

The screenshot shows the 'Interactive Workbench - Comments' interface. The top menu bar includes File, Edit, View, Generate, Categories, Tools, and Help. A toolbar below the menu contains icons for Build, Extend, Score, and Display. The main workspace is divided into several panes:

- Category Bar:** Displays a hierarchical tree of categories. The root node is 'All Documents', which branches into 'Uncategorized' and 'No concepts extracted'. Under 'No concepts extracted', there are nodes for 'music', 'memory device', 'consumer electronics', 'electronics', 'songs', 'color', and 'clothing and dress'. The total number of documents is 904.
- Category Web:** A table showing the selection percentage and count for various categories. The top row is 'electronic' at 100.0% (115 docs). Other categories include music (7.8%, 9 docs), players (1.7%, 2 docs), internet (1.7%, 2 docs), design (0.9%, 1 doc), listening (0.9%, 1 doc), plays (1.7%, 2 docs), memory d (3.5%, 4 docs), and earphone (1.7%, 2 docs).
- Concept:** A table listing 539 concepts. The columns include Concept, In, Global, Docs, and Type. Notable concepts include 'battery' (117 (5%) docs, type <Performance>), 'small' (69 (3%) docs, type <Contextual>), 'music' (69 (3%) docs, type <Features>), 'like' (61 (3%) docs, type <Positive>), 'easy to use' (52 (2%) docs, type <Positive>), 'sound' (50 (2%) docs, type <Features>), 'nothing' (46 (2%) docs, type <Uncertain>), 'good' (46 (2%) docs, type <Positive>), 'portable' (44 (2%) docs, type <Positive>), and 'expensive' (43 (2%) docs, type <NegativeBud...>).
- Comments (115):** A list of 6 comments, each associated with a category. The comments are numbered 1 through 6 and include:
 - 1: Battery life, Portability, Accessories, Style. Category: electronics/battery
 - 2: batteries last a long time. Category: electronics/battery
 - 3: Long battery life. Also when you turn it on it picks up where you left off. Category: electronics/battery
 - 4: Easy to organise my music files. Good battery life. Category: electronics/battery, music/music files
 - 5: It's lightweight, the battery life is very good, cool design. Category: electronics/battery, design
 - 6: I love it. I can listen to old heavy metal and new metal and alternative. Its batteries last for days. Category: electronics/battery, listening, music/music genres
- Categories:** A table showing the count of categories. It lists 30 categories with 52 sub-categories.

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 if one was not.

Table (69 fields, 904 records)

| | Category_holding | Category_plays | Category_time | Category_feature | Category_commuting | Category_work | Category_memory device | Category_user interface | Category_listening |
|----|------------------|----------------|---------------|------------------|--------------------|---------------|------------------------|-------------------------|--------------------|
| 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 | T | F | F |
| 19 | F | F | F | F | F | F | F | 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 |

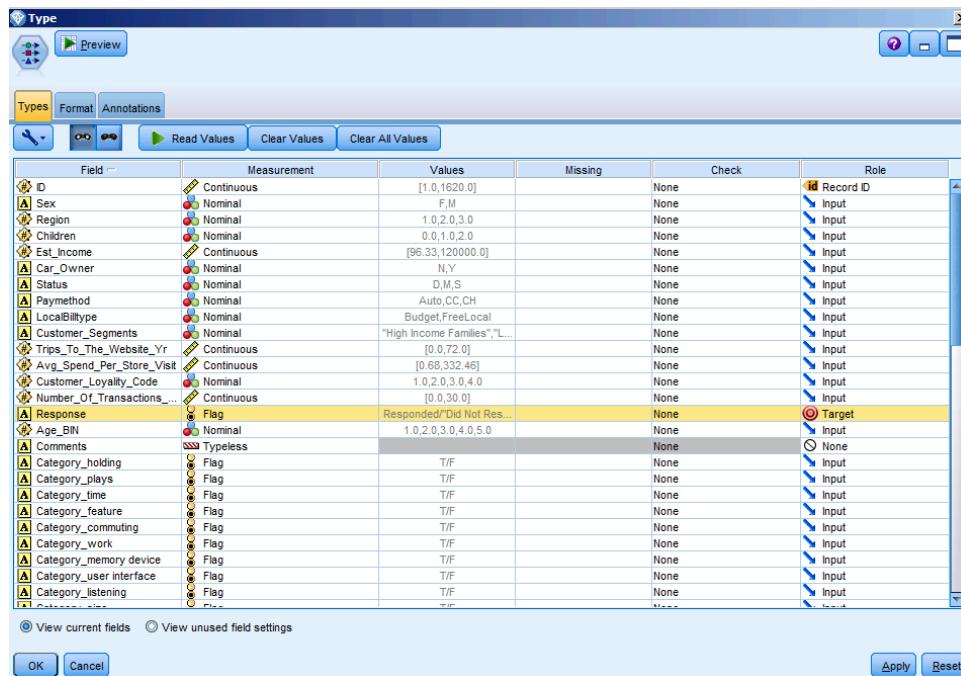
Annotations

OK

This screenshot shows the Output palette with a table view. The table has 69 columns and 904 rows. The first few columns are labeled with category names: Category_holding, Category_plays, Category_time, Category_feature, Category_commuting, Category_work, Category_memory device, Category_user interface, and Category_listening. Each row contains a series of 'F' and 'T' characters, representing binary values for each category across the 904 records. The 'Annotations' tab is selected at the bottom of the palette.

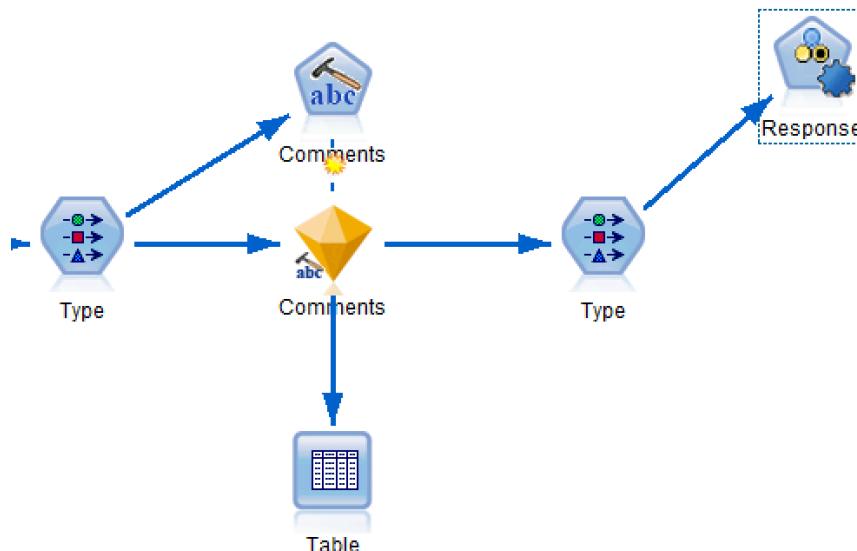
11. To properly define the newly created categories, we need to use a Type node. Add a Type node to the stream, connecting it to the modeling node.

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.

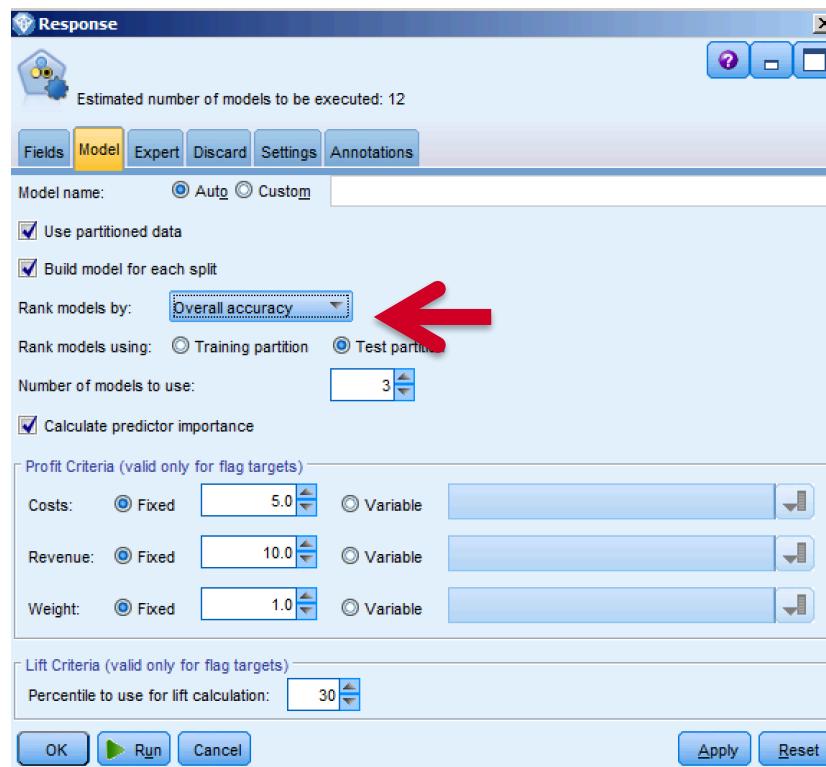


12. To understand the impact these newly created Fields might have on our ability to predict campaign response, we will build another model.

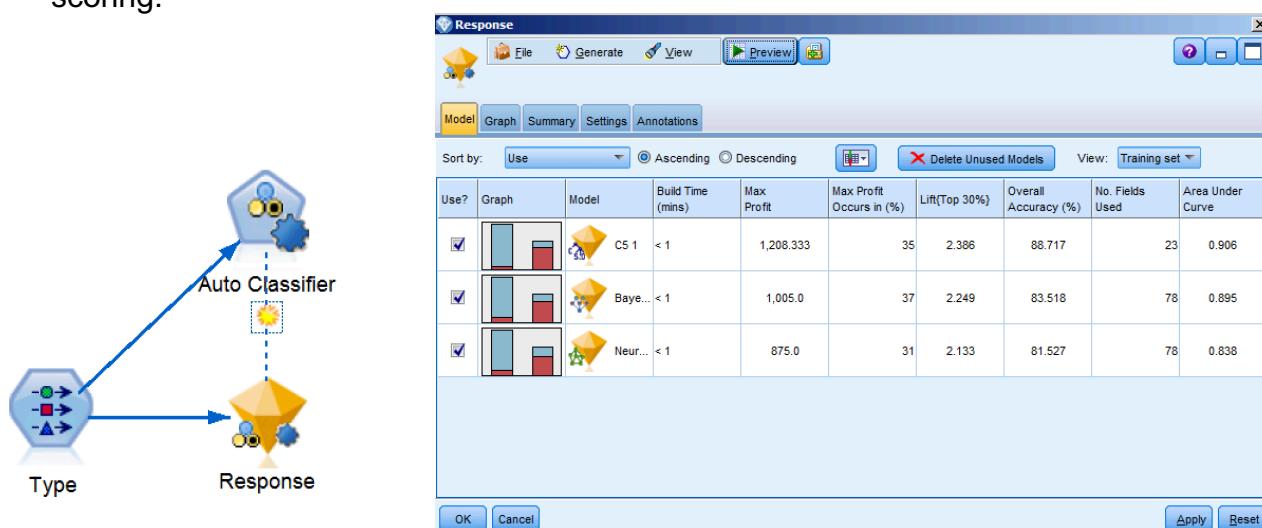
Select the Auto Classifier node from the Modeling palette and connect it to the Type node. The Auto Classifier node creates and compares a number of different models for binary outcomes (yes or no, responded or did not respond, and so on), allowing you to choose the best approach for a given analysis.



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.

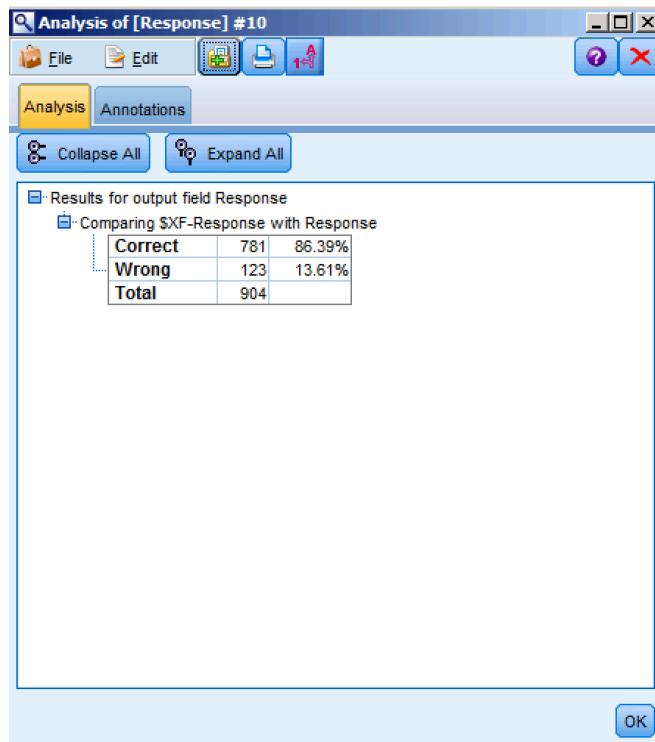


The Auto Classifier model is automatically generated and added to our canvas. When the Auto Classifier is executed, the node estimates candidate models for every possible combination of options, ranks each candidate model based on the measure you specify, and saves the best models in a composite automated 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.



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%.



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

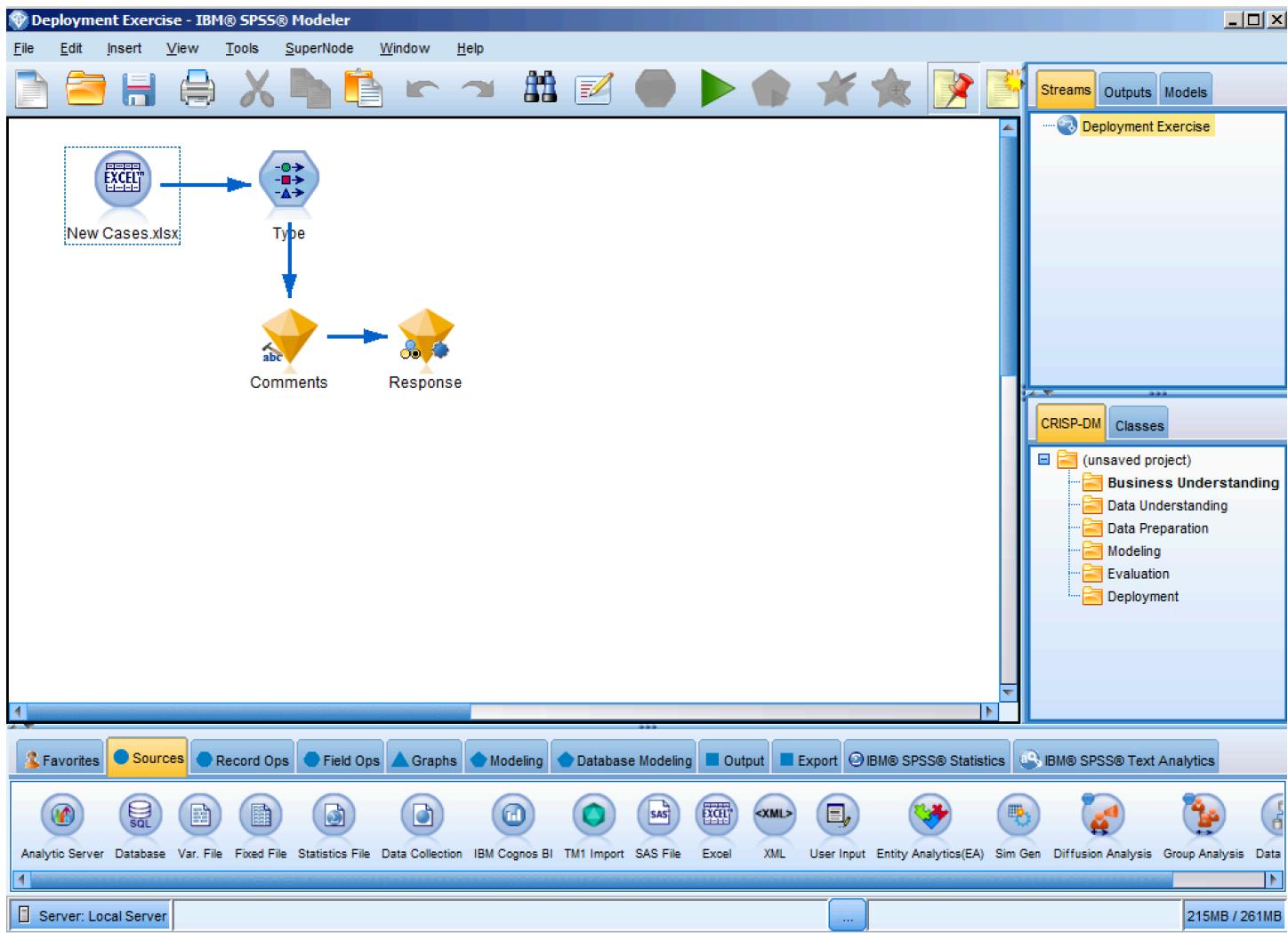
94% reduction in manual processes, allowing the company to offer services at a more competitive price

"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

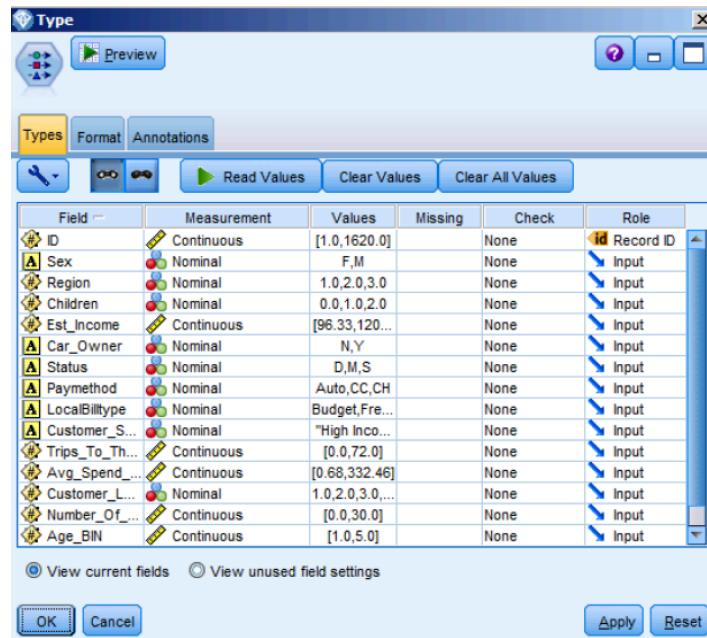
- To start, open the stream labeled “Deployment Exercise.str” from the workshop directory. In IBM SPSS Modeler, click on File, Open stream, and then navigate to C:\Modeler\Workshop\Deployment\Deployment Exercise.str.

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



For this exercise, 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 includes 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.



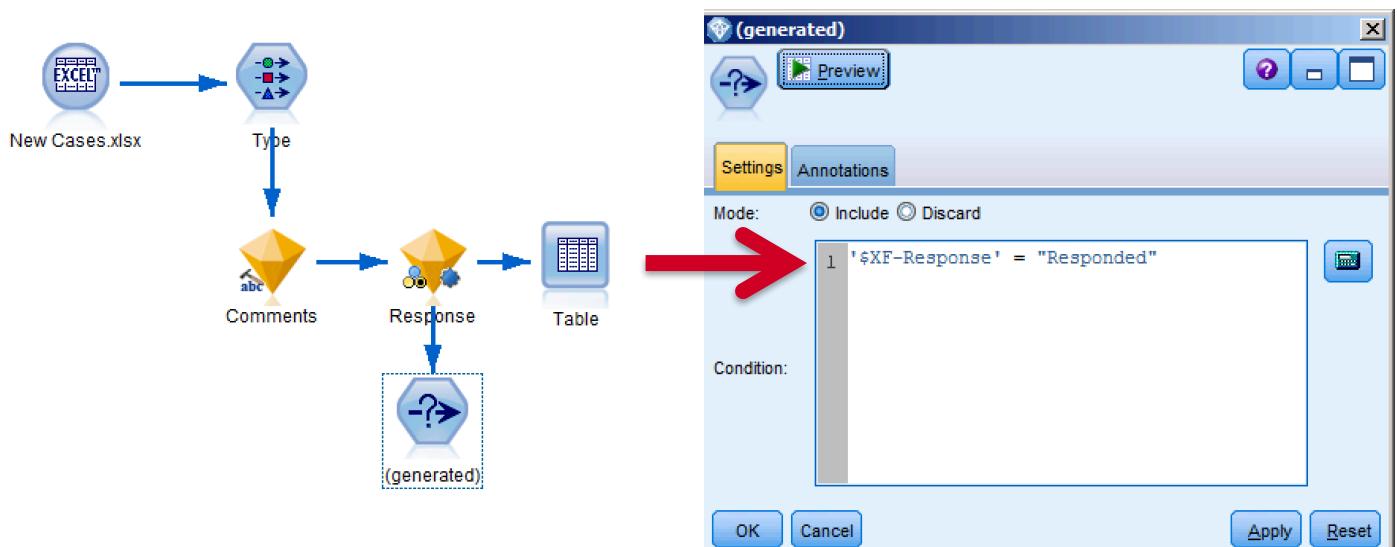
- From the Output palette, add a Table node to the canvas, connect it to the Response model node, and select Run.

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.

| /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 |

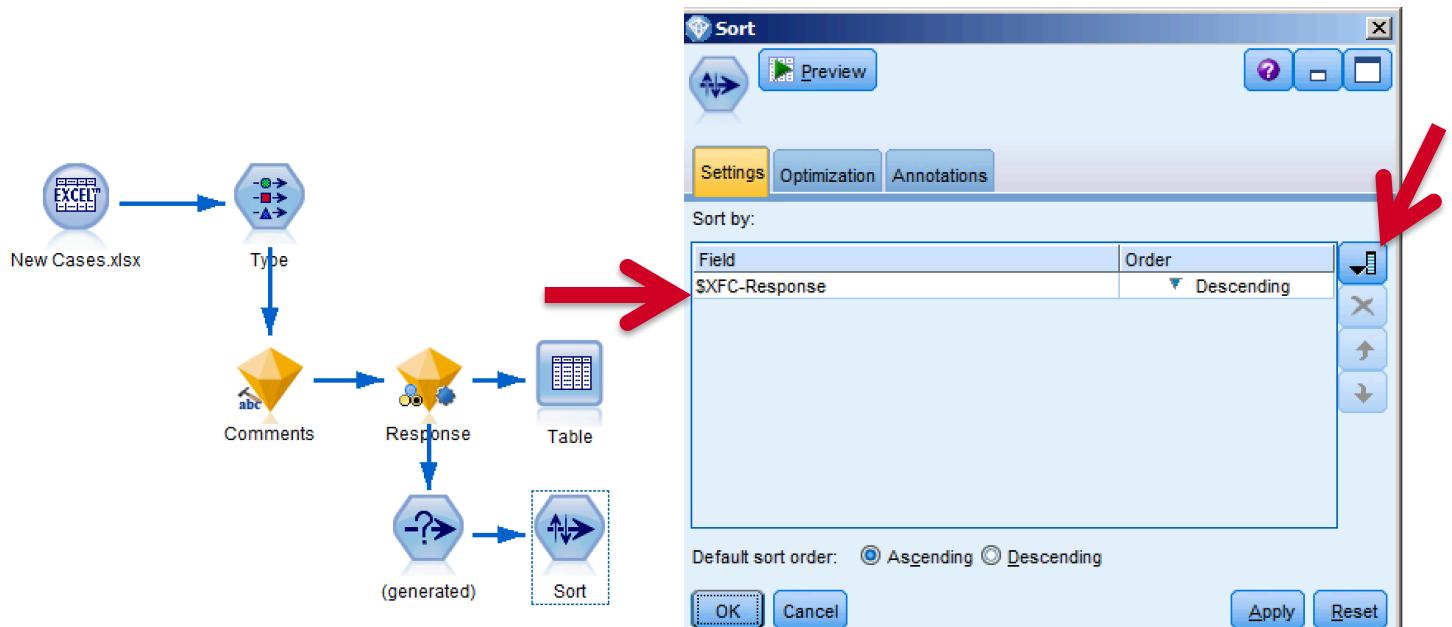
- Highlight one of the cells with the value of Responded. From the drop-down menu, select Generate from the tool bar and then choose Select Node ("Or"). This will generate a Select node on canvas. Alternatively, you can add a Select node form the Record Ops palette (not shown), and connect it to the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



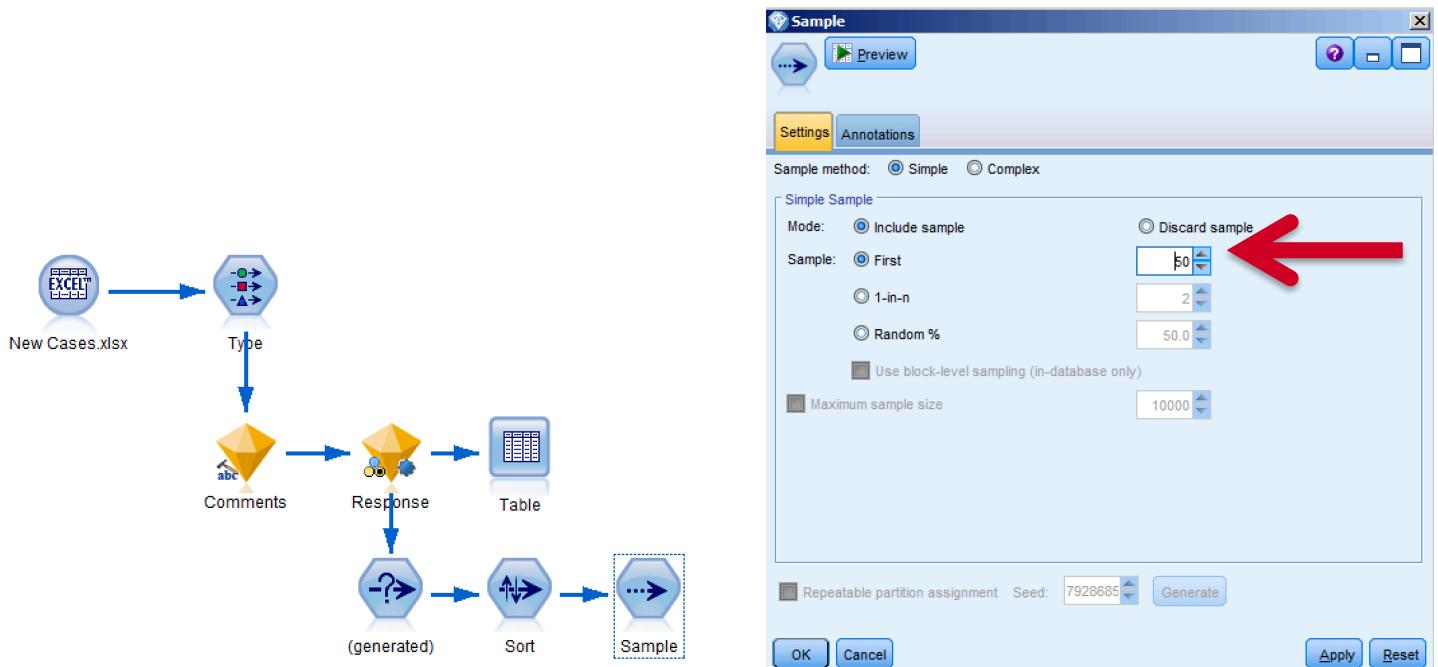
- From the Record Ops palette, select a Sort node, and connect it to the Generated node.

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.



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

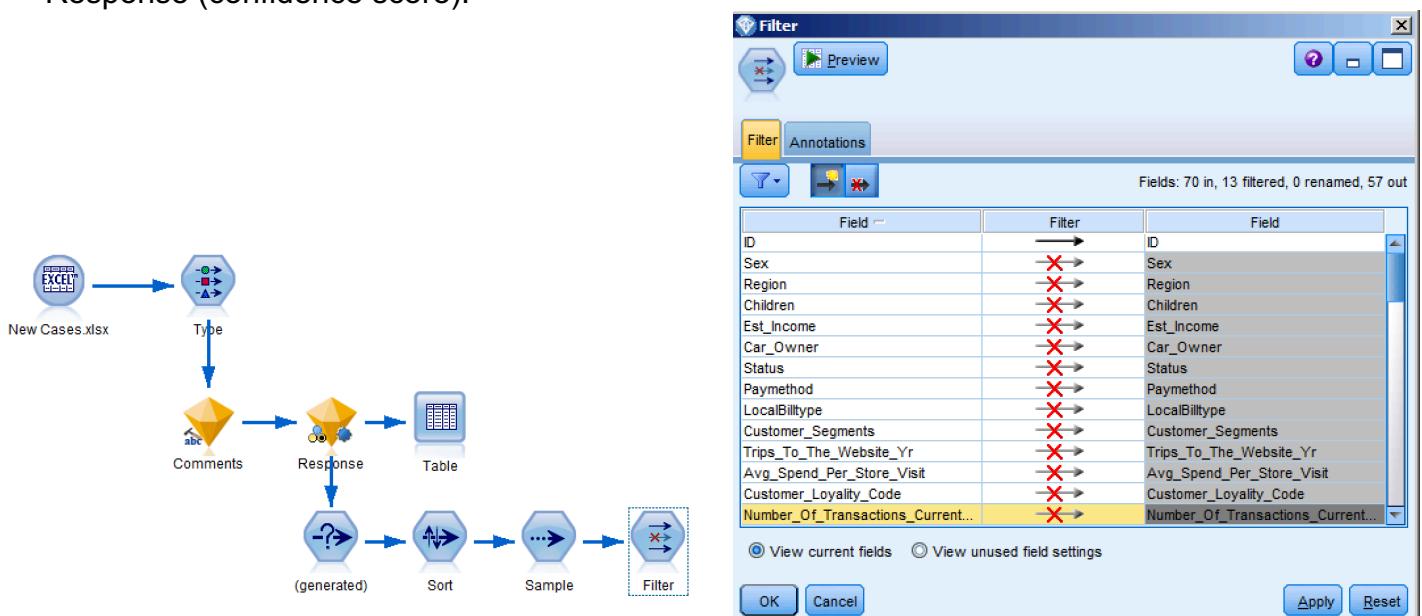
Double-click the Sample node to edit the settings, sampling the first 50 records. Click OK.



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.

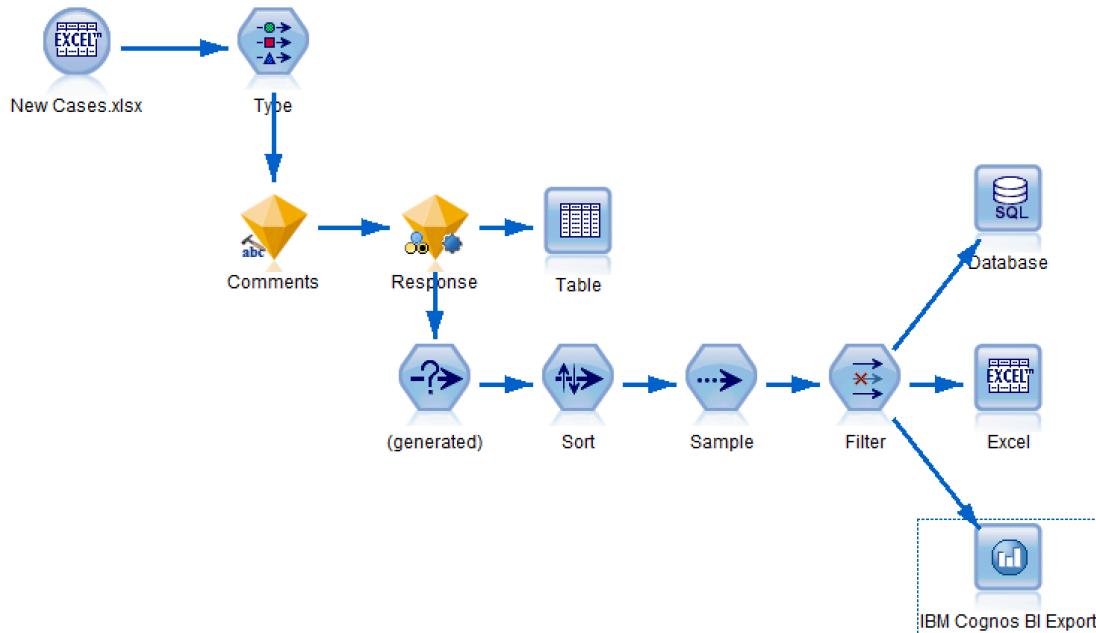
From the Fields Ops palette, select a Filter node, and connect it to the Sample node.

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).



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.

The screenshot shows a Microsoft Excel spreadsheet titled "Target List.xlsx". The spreadsheet has a single sheet named "Sheet1". The data is organized into columns A and B. Column A is labeled "ID" and contains numerical values from 1 to 15. Column B is labeled "\$XF-Response" and contains corresponding response percentages. The data is as follows:

| | ID | \$XF-Response |
|----|------|---------------------|
| 1 | 308 | Responded 92.96% |
| 2 | 444 | Responded 92.81% |
| 3 | 1331 | Responded 92.57% |
| 4 | 494 | Responded 91.86% |
| 5 | 801 | Responded 91.08% |
| 6 | 458 | Responded 90.38% |
| 7 | 1132 | Responded 90.01% |
| 8 | 943 | Responded 89.83% |
| 9 | 1318 | Responded 89.55% |
| 10 | 1470 | Responded 89.43% |
| 11 | 65 | Responded 88.35% |
| 12 | 577 | Responded 88.34% |
| 13 | 548 | Responded 87.16% |
| 14 | 778 | Responded 86.91% |
| 15 | | |

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.

© Copyright IBM Corporation 2017

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

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.