SPSS Modeler Tutorial 2

– The Market Basket Project

Data Warehousing and Data Mining

March 2014

2. The Market Basket Project

Briefing: This example deals with fictitious data describing the contents of supermarket baskets (that is, collections of items bought together) plus the associated personal data of the purchaser, which might be acquired through a loyalty card scheme. The goal is to discover groups of customers who buy similar products and can be characterized demographically, such as by age, income, and so on.

This example illustrates two phases of data mining:

- Association rule modeling and a web display revealing links between items purchased
- C5.0 rule induction profiling the purchasers of identified product groups

Note: This application does not make direct use of predictive modeling, so there are no accuracy measurements for the resulting models and no associated training/test distinction in the data mining process.

2.1 Accessing the data

Open the SPSS Modeler by going to the Start menu → All Programs → IBM SPSS Modeler 15.0 → IBM SPSS Modeler 15.0. Select "Open an existing project" and double-click on "More files…". In the Open dialog window, go to the path of "N:\DWDM\SPSSModeler\Demos" and double-click on the "bask.cpj" file to open it.

In this project, we need to use the data file "BASKETS1n".

1. Select the "Var.File" node listed in the "Sources" tab from the "Module Panel", and add it to the "Main Panel".
2. Double click the "Var.File" node in the "Main Panel" to open its property window, and Click the "…" button next to the "File" field. In the "Open" dialog window, select to open the "BASKETS1n" file that contains records of basket information (Figure 1). The BASKETS1n file contains records for 18 attributes, termed "cardid", "value", "pmethod", "sex", "homeown", "income", "age", "fruitveg", "freshmeat", "dairy", "cannedveg", "cannedmeat", "frozenmeal", "beer", "wine", "softdrink", "fish", and "confectionery".
3. Click "OK" to close the "Var.File" property window.

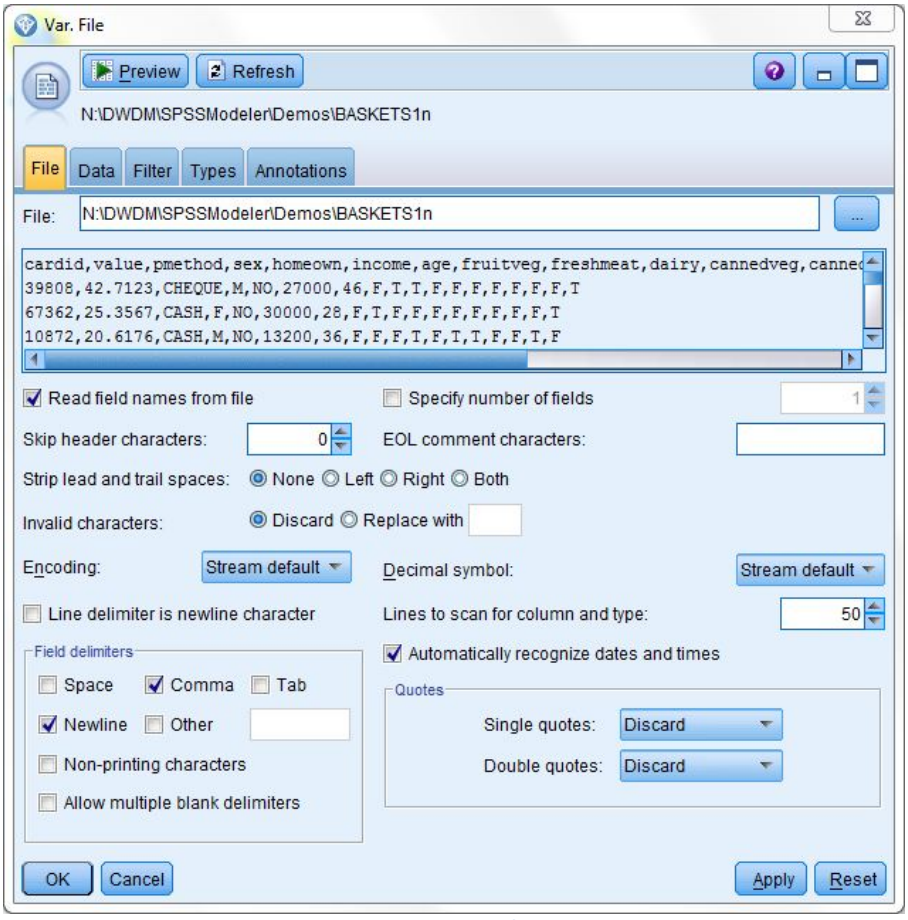


Figure 1: BASKETS1n File Property

2.2 Find and display associations between data attributes.

1. Select the "Type" node listed in the "Field Ops" tab from the "Module Panel", and add it to the "Main Panel".
2. Establish a link between the "BASKET1n" node and the "Type" node by right-clicking on the "BASKET1n" node and select the "Connect…" option, then left-clicking on the "Type" node (Figure 2).

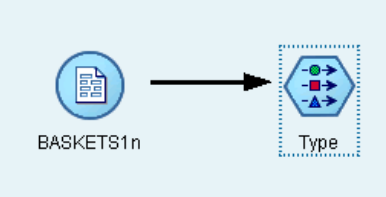
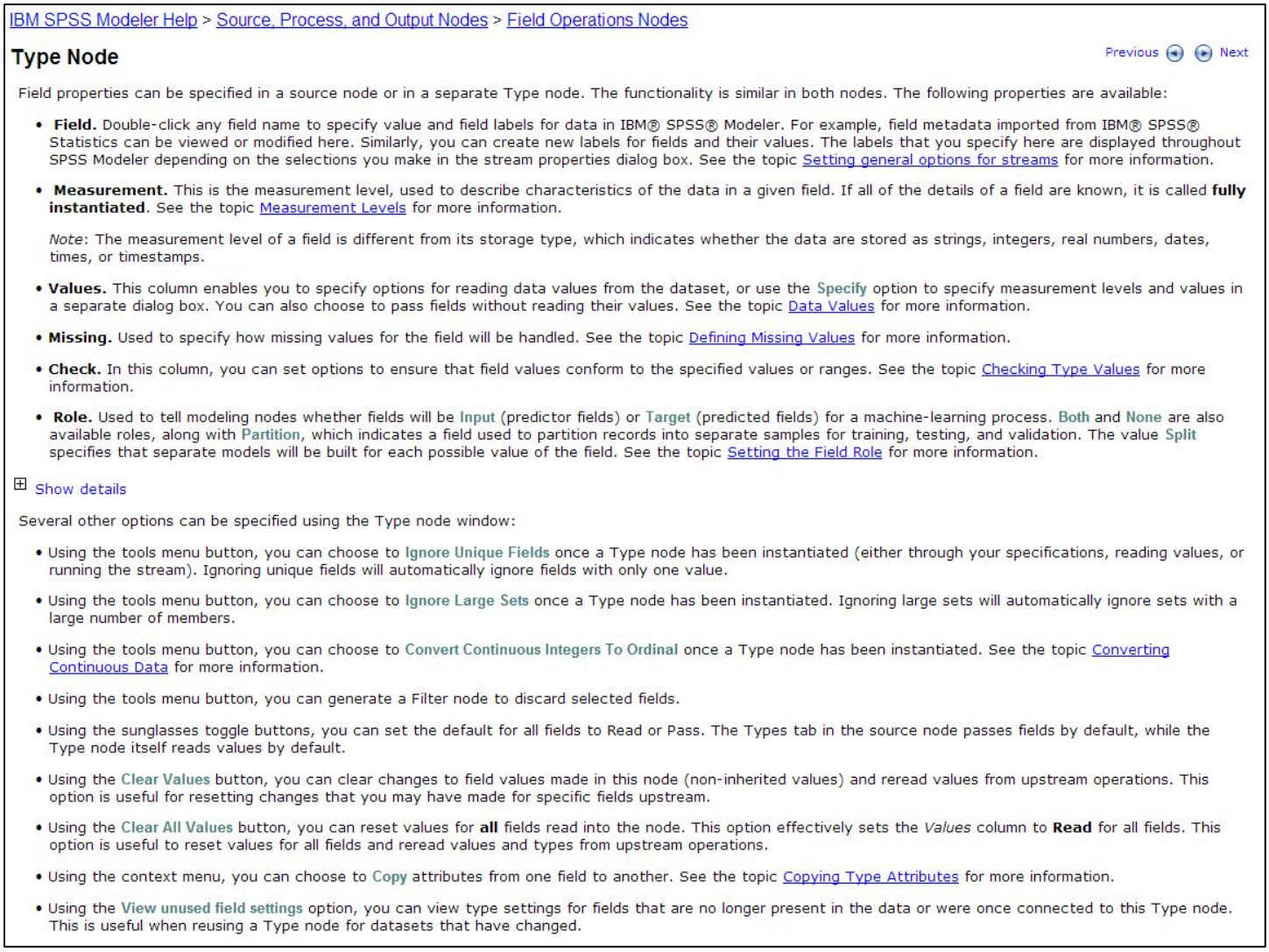


Figure 2: Link between BASKET1n and Type Nodes

3. Double-click the "Type" node to open its property window. The "Type" node provides a way to modify the property of data attributes in the source node it connects to. Full detail of the "Type" node can be found in the Help file by clicking on the Help button and selecting "Type Node" (Figure 3).

Note: You can click the "Read Values" button to detect value ranges for data attributes in the data source.



IBM SPSS Modeler Help > Source, Process, and Output Nodes > Field Operations Nodes

Type Node

Previous Next

Field properties can be specified in a source node or in a separate Type node. The functionality is similar in both nodes. The following properties are available:

- **Field.** Double-click any field name to specify value and field labels for data in IBM® SPSS® Modeler. For example, field metadata imported from IBM® SPSS® Statistics can be viewed or modified here. Similarly, you can create new labels for fields and their values. The labels that you specify here are displayed throughout SPSS Modeler depending on the selections you make in the stream properties dialog box. See the topic Setting general options for streams for more information.
- **Measurement.** This is the measurement level, used to describe characteristics of the data in a given field. If all of the details of a field are known, it is called **fully instantiated**. See the topic Measurement Levels for more information.

  Note: The measurement level of a field is different from its storage type, which indicates whether the data are stored as strings, integers, real numbers, dates, times, or timestamps.
- **Values.** This column enables you to specify options for reading data values from the dataset, or use the Specify option to specify measurement levels and values in a separate dialog box. You can also choose to pass fields without reading their values. See the topic Data Values for more information.
- **Missing.** Used to specify how missing values for the field will be handled. See the topic Defining Missing Values for more information.
- **Check.** In this column, you can set options to ensure that field values conform to the specified values or ranges. See the topic Checking Type Values for more information.
- **Role.** Used to tell modeling nodes whether fields will be Input (predictor fields) or Target (predicted fields) for a machine-learning process. Both and None are also available roles, along with Partition, which indicates a field used to partition records into separate samples for training, testing, and validation. The value Split specifies that separate models will be built for each possible value of the field. See the topic Setting the Field Role for more information.

⊞ Show details

Several other options can be specified using the Type node window:

- Using the tools menu button, you can choose to Ignore Unique Fields once a Type node has been instantiated (either through your specifications, reading values, or running the stream). Ignoring unique fields will automatically ignore fields with only one value.
- Using the tools menu button, you can choose to Ignore Large Sets once a Type node has been instantiated. Ignoring large sets will automatically ignore sets with a large number of members.
- Using the tools menu button, you can choose to Convert Continuous Integers To Ordinal once a Type node has been instantiated. See the topic Converting Continuous Data for more information.
- Using the tools menu button, you can generate a Filter node to discard selected fields.
- Using the sunglasses toggle buttons, you can set the default for all fields to Read or Pass. The Types tab in the source node passes fields by default, while the Type node itself reads values by default.
- Using the Clear Values button, you can clear changes to field values made in this node (non-inherited values) and reread values from upstream operations. This option is useful for resetting changes that you may have made for specific fields upstream.
- Using the Clear All Values button, you can reset values for **all** fields read into the node. This option effectively sets the *Values* column to **Read** for all fields. This option is useful to reset values for all fields and reread values and types from upstream operations.
- Using the context menu, you can choose to Copy attributes from one field to another. See the topic Copying Type Attributes for more information.
- Using the View unused field settings option, you can view type settings for fields that are no longer present in the data or were once connected to this Type node. This is useful when reusing a Type node for datasets that have changed.

Figure 3: Description of Labels in the "Type" node

4. Modify the properties of attributes as in Figure 4.
   a. Set the “Measurement” property of “cardid” to “Typeless”. This is because that each loyalty card ID occurs only once in the dataset can therefore be of no use in modeling.
   b. Set the “Measurement” property of “Sex” to “Nominal”. This is to ensure that the Apriori modeling algorithm will not treat “sex” as a flag.
   c. Set the “Role” property to “None” for “cardid”, “value”, “pmethod”, “sex”, “homeown”, “income”, and “age”.
   d. Set the “Role” property to “Both” for the remaining attributes.
   e. Click “OK” to close properties window
5. Select the “Apriori” node listed in the “Modelling” tab from the “Module Panel”, and add it to the “Main Panel”. Apriori node discovers association rules in the data.
6. Connect the “Apriori” node to the “Type” node in the “Main Panel” (Figure 5).
7. Double click the “Apriori” node to open its property window.
8. Click “Run”. It creates a new model. Double-click this model and you can observe a table that displays detected associations between data attributes, which roles are set to “Both” in step 4, should appear as in Figure 6. These rules show a variety of associations between frozen meals, canned vegetables, and beer.

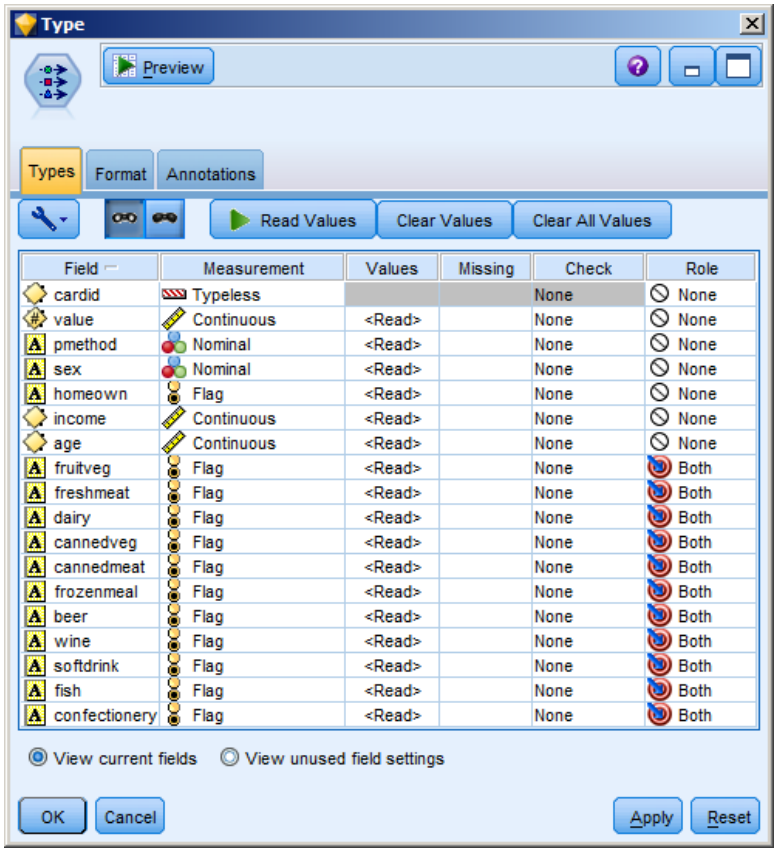


Figure 4: Modified Properties in the "Type" node

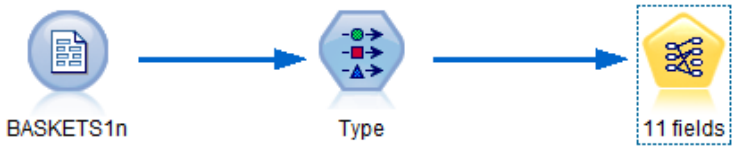


Figure 5: Connections between BASKETS1n, Type, and Apriori nodes

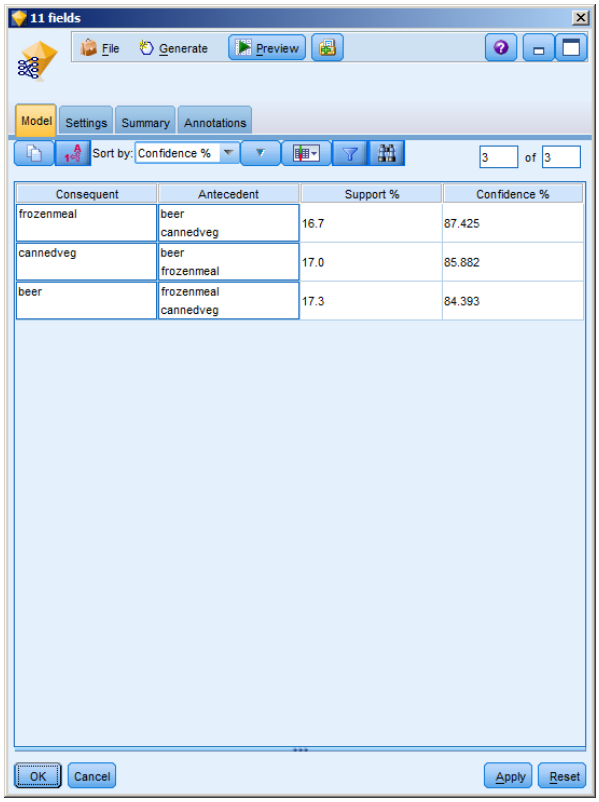


Figure 6: Associations between Data Attributes

9. Select the "Web" node listed in the "Graphs" tab from the "Module Panel", and add it to the "Main Panel".
10. Connect the "Web" node to the "Type" node to have a visual view of how different data attributes are associated as in Figure 7.

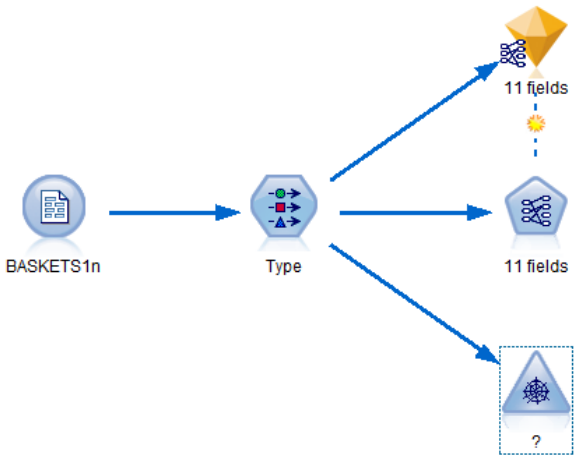


Figure 7: Connection between Web and Type Nodes

11. Double click the "Web" node to open its property window.
12. Using the Select Fields drop down menu, select "fruitveg", "freshmeat", "dairy", "cannedveg", "cannedmeat", "frozenmeal", "beer", "wine", "softdrink", "fish", and "confectionery" for the "Fields", and tick "Show true flags only" (Figure 8).

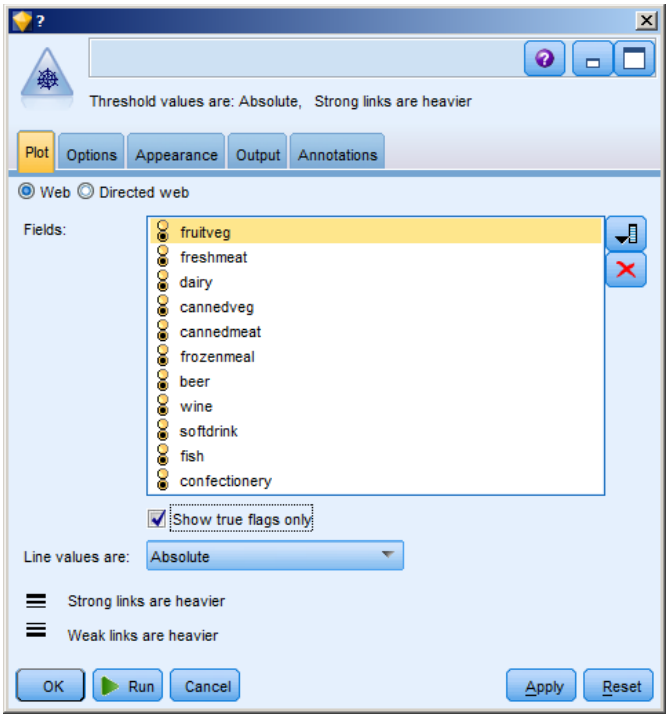


Figure 8: Property Window of Web node

13. Click "Run" and a graphical display of associations between data attributes should be generated as Figure 9. Your result may look different from Figure 9. This is because the threshold used, which can be set using the scroll bar at the bottom of the window.

    We can observe that three groups of customers stand out

    - Those who buy fish and fruits and vegetables, who might be called Healthy eaters
    - Those who buy wine and confectionery
    - Those who buy beer, frozen meals, and canned vegetables (Beer, beans, and pizza)

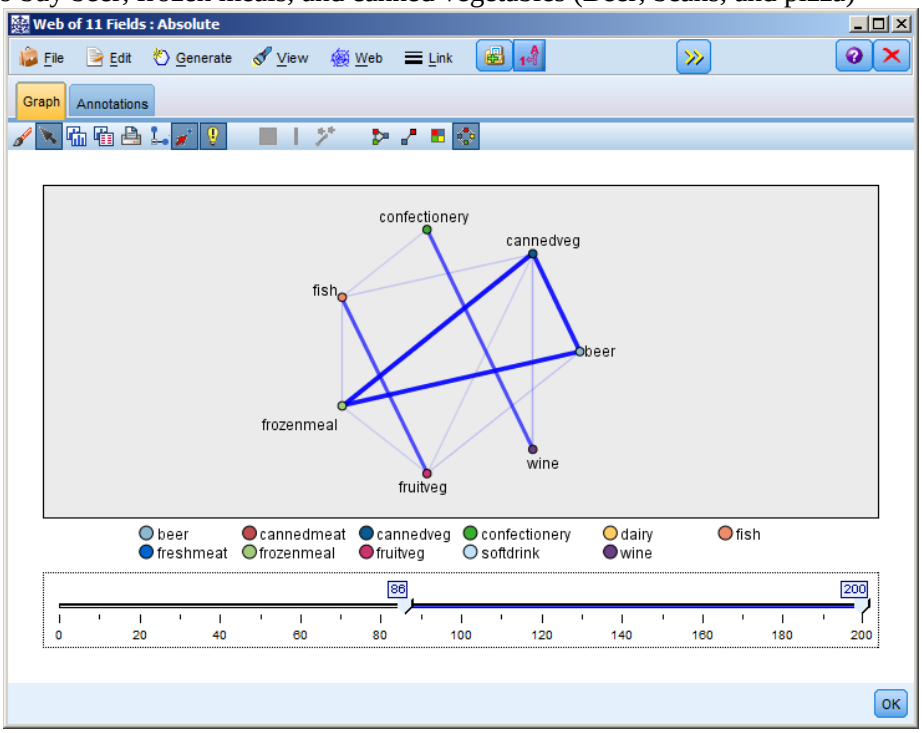


Figure 9: Result of Web Node

2.3 Profiling the Customer Groups

You have now identified three groups of customers based on the types of products they buy, but you would also like to know who these customers are, their demographic profile. This can be achieved by tagging each customer with a flag for each of these groups and using rule induction (C5.0) to build rule-based profiles of these flags.

1. You must derive a flag for each group. This can be automatically generated using the web display that you just created. Using the right mouse button, click the link between “fruitveg” and “fish” and select “Generate Derive Node for Link”. A new node should appear in the “Main Panel”.

2. Double click the newly generated node to open its property window, and change the “Derive field” to "Healthy” (Figure 10).

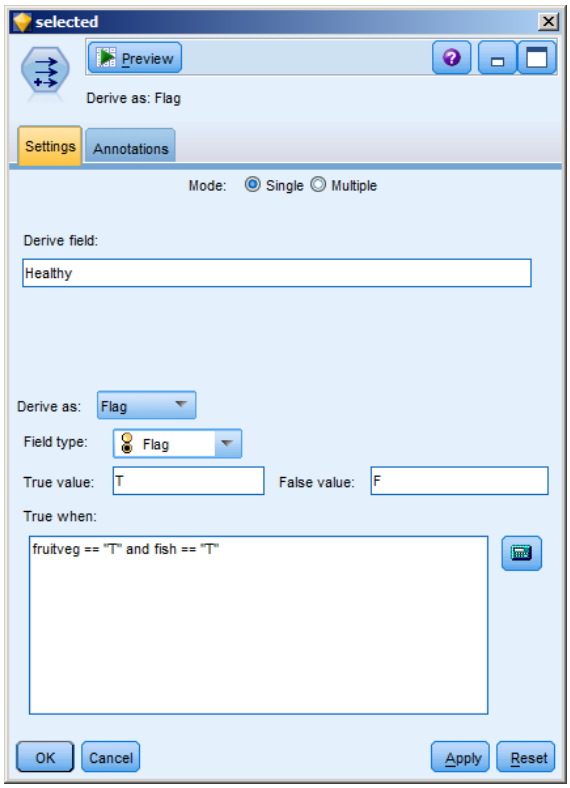


Figure 10: The Healthy Node

3. Repeat Step 1 and 2 for the link between “Wine” and “Confectionery”, and rename the derived node to “WineChocs”.
4. Repeat Step 1 and 2 for the links between “cannedveg”, “beer”, and “frozenmeal”. To derive a node from multiple links, you need to
    a. Goto the “interaction” mode, by selecting “Interactions” from the “View” menu.
    b. Select the “magic wand” – it appears as a magic wand icon with two red starts on the Graph menu.
    c. Use the magic wand to draw a line crossing the first link you want to select (Be careful, if you draw a line across multiple links, they will all be selected).
    d. While holding the “Shift” key, repeat for each other link you want to select.
    e. Then select “Devive Node (“And”) option from the “Generate” menu (Figure 11).

    A new node will be generated in the “Main Panel”. Rename it as “beer_beans_pizza”.

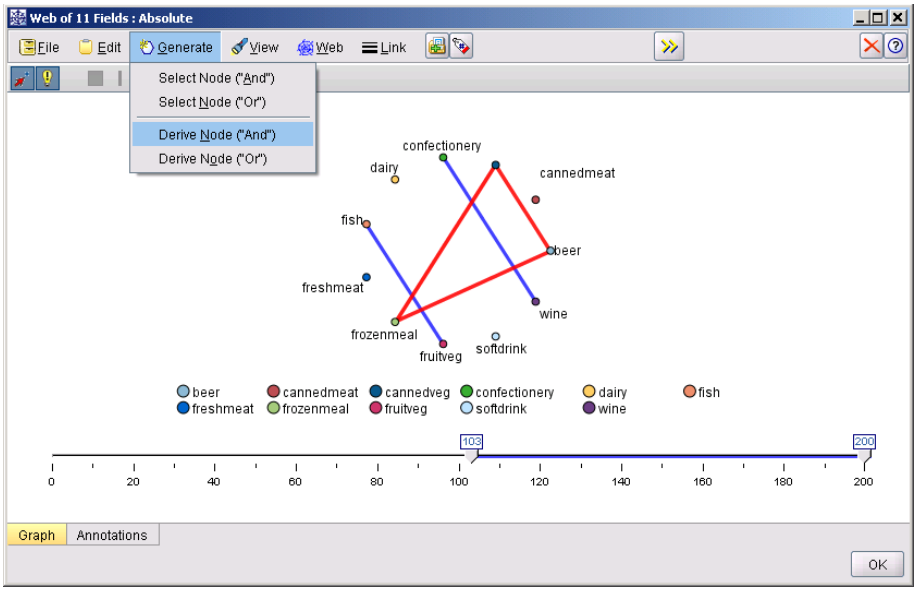


Figure 11: Derive a Node from Multiple Links

5. To profile these customer groups, connect the existing "Type" node to these three newly generated nodes in series, and then attach another "Type" node. (Figure 12).

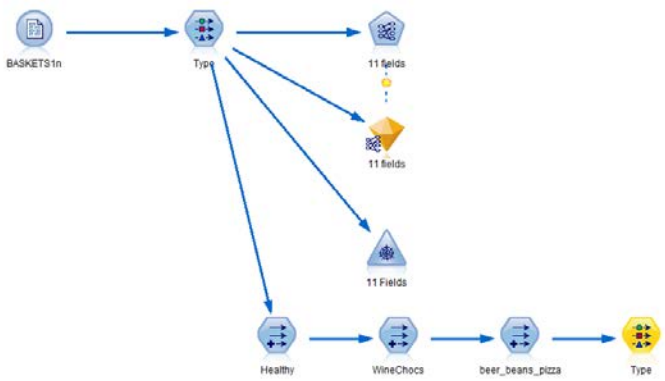


Figure 12: Connections for customer profiling

6. Double click the new "Type" node to open its property window.
    a. Set "Role" for "value", "pmethod", "sex", "homeown", "income", and "age" to "Input";
    b. Set "Role" for a customer group, which is <u>one of</u> "Healthy", "WineChocs", and "beer_beans_pizza", to "Target".
    c. Set "Role" for the remaining data attributes to "None" (Figure 13).
7. Click OK to close the property window.

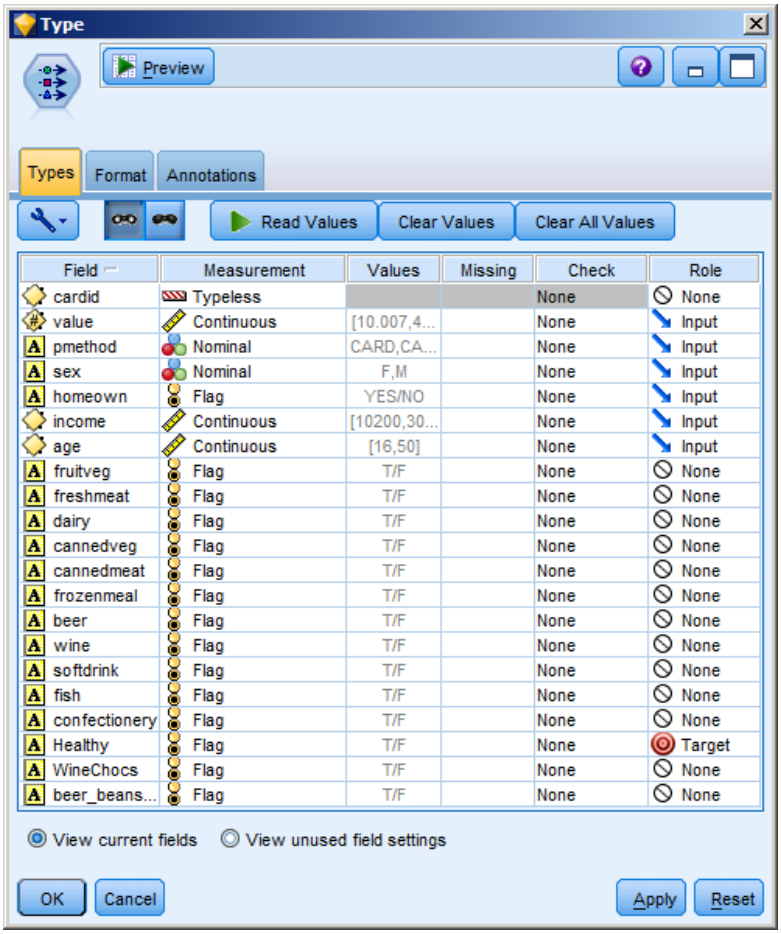


Figure 13: Modified Type Node

8. Select the "C5.0" node listed in the "Modelling" tab from the "Module Panel", and add it to the "Main Panel".
9. Double click the "C5.0" node to open its property window.
10. Set the "Output type" to "Rule set" (Figure 14).

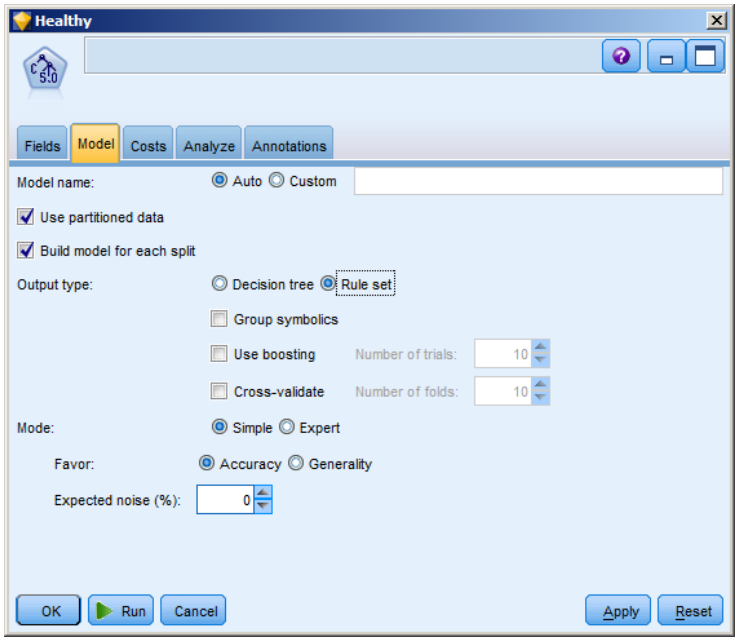


Figure 14: Property Window of C5.0 Node

11. Click "Run", and a new model will be generated in the "Current Working Space" area.

12. Double click on the new model.

13. The result shows a clear demographic profile for this customer group (Figure 15).

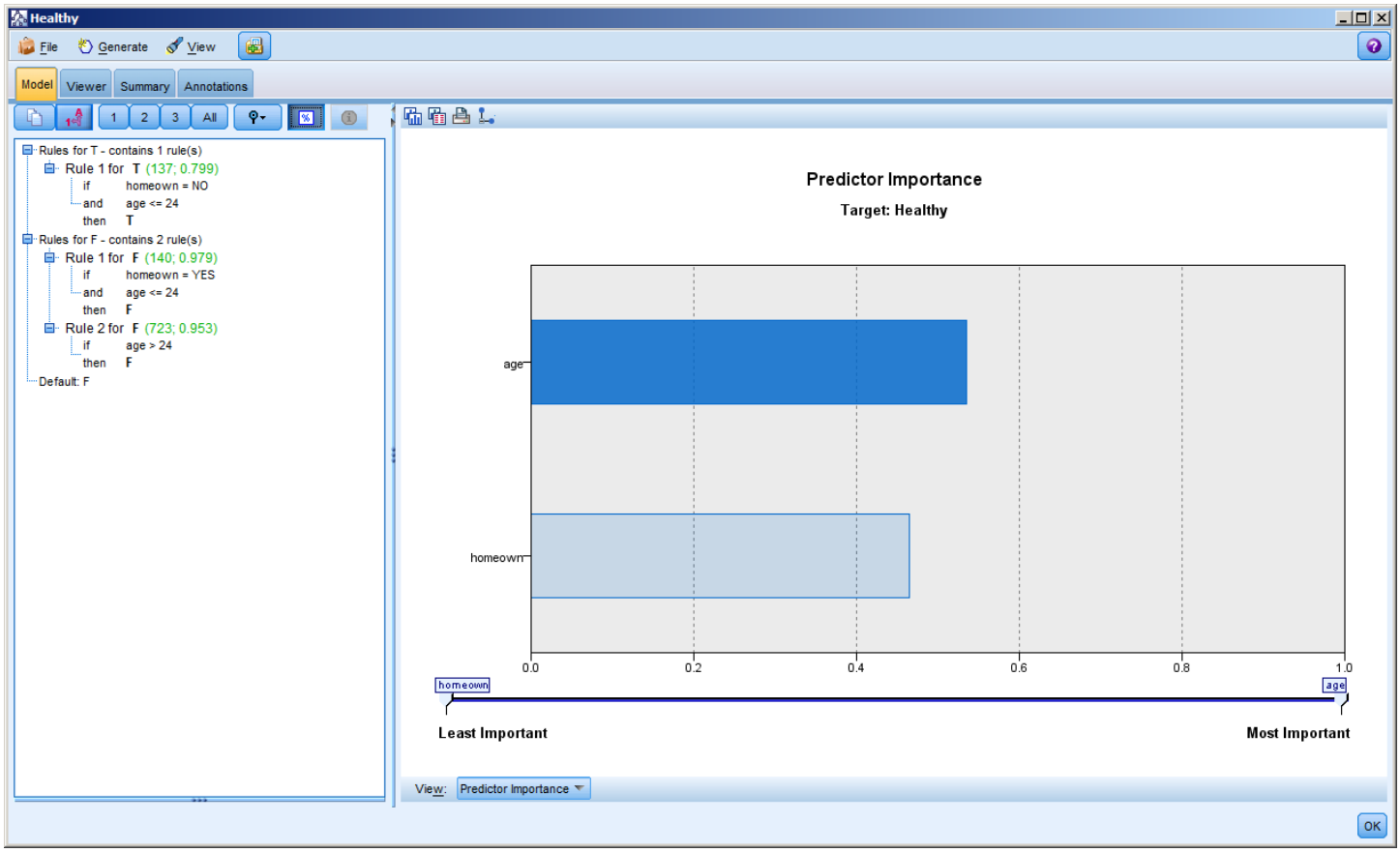


Figure 15: Demographic Profile for beer_beans_pizza Customer Group

The same method can be applied to the other customer group flags by selecting them as the output in the second Type node. A wider range of alternative profiles can be generated by using Apriori instead of C5.0 in this context; Apriori can also be used to profile all of the customer group flags simultaneously because it is not restricted to a single output field.

End of Tutorial 2