



# Data Mining with SQL Server Data Tools

**Data mining tasks include classification (directed/supervised) models as well as (undirected/unsupervised) models of association analysis and clustering.**

## Data Mining

Data mining has many definitions and may be called by other names such as knowledge discovery. It is generally considered to be a part of the umbrella of tasks, tools, techniques etc. within business Intelligence (BI). Many corporate managers consider BI to be the heart of all the processes that support decision making at all levels. A definition of data mining typically includes large datasets, discovering previously unknown knowledge and patterns and that this knowledge is actionable. That what is discovered is not trivial but can be usefully applied. BI and its Data Mining component are receiving considerable attention and fanfare as companies utilize BI for competitive advantage.

Different authors may address the data mining tasks slightly different from each other but the following terminology provides a helpful and useful basis for discussing data mining. The data mining tasks are:

- Description
- Estimation
- Classification
- Prediction
- Association Analysis
- Clustering

**Description**—used descriptive statistics to better understand and profile areas of interest. Thus a variety of well known statistical tools and methods are used for this task—including frequency charts and other graphical output, measures of central tendency and variation.

### Data Mining Tasks with a Target or Dependent Variable

**Estimation, classification** and **prediction** are data mining tasks that have a target (dependent) variable. Sometimes these, are referred to as predictive analysis; however, many authors reserve the term Prediction to use of models for the future. The terms **supervised** and **directed** apply to these data mining tasks. **Estimation** data mining tasks have an interval level dependent target variable whereas **classification** data mining tasks have a categorical (symbolic) target variable. An example of an estimation data mining task would be estimating family income based on a number of attributes; whereas a model to place families into the three income brackets of Low, Medium or High would be an example of a classification data mining task. Thus, the difference between the two tasks is the type of target variable.

When either an estimation data mining task or classification task is used to predict future outcomes, the data mining task becomes one of **Prediction**. Again, estimation and classification are referred to as predictive models because that would be the typical application of models built for these data mining tasks.

In summary, the most important concept is that estimation and classification data mining tasks require a target variable. However, the difference lies in the data type of the target variable.

**Data Mining Algorithms for Directed/Supervised Data Mining Tasks**—**linear regression** models are the most common data mining algorithms for **estimation** data mining tasks. Of course, linear regression is a very well known and familiar technique. A number of data mining algorithms can be used for **classification** data mining tasks including **logistic regression, decision trees, neural networks, memory based reasoning (k-nearest neighbor), and Naïve Bayes**.

## Data Mining Tasks without a Target or Dependent Variable

**Association Analysis** and **Clustering** are data mining tasks that do not have a target (dependent) variable. Affinity analysis is another term that refers to association analysis and is typically used for market basket analysis (MBA) although association analysis can be used for other areas of study. MBA is essentially analyzing what purchases tend to be purchased together—that is what items tend to have an affinity with other items. **Clustering**, having no target variable, algorithms attempt to put records into groups based on the record's attributes. The critical concept is that of similarity—those within a cluster are very similar to each other and not similar with those in another cluster.

**Note**—because these data mining tasks do not have a target variable, their corresponding models cannot be used for prediction. Thus, they are many times exploratory in nature and their results can be used downstream in predictive models.

### Data Mining Examples in this Tutorial

The data mining tasks included in this tutorial are the directed/supervised data mining task of classification (Prediction) and the undirected/unsupervised data mining tasks of association analysis and clustering. Many users already have a good linear regression background so estimation with linear regression is not being illustrated. Three data mining algorithms for the classification data mining tasks will be illustrated and compared: **Decision Trees**, **Logistic Regression**, and **Neural Networks**. Recall that classification has a categorical target variable.

Association analysis and clustering are the undirected/unsupervised data mining tasks illustrated in this tutorial. The clustering algorithm is **k-means**.

### Data mining overview summary

Data mining tasks	Target Variable	Typical Data Mining Algorithm(s)
Description	No	Statistics, including descriptive, & visualization
Estimation	Yes Interval Numeric	Linear Regression
Classification	Yes Categorical	Logistic Regression, Decision Trees, Neural Networks, Memory Based Reasoning, Naïve Bayes
Prediction	Yes	Estimation and Classification models for prediction
Association Analysis	No	Affinity Analysis (Market Basket Analysis)
Clustering	No	k-means, Kohonen Self Organizing Maps (SOM)

## Data Mining Example using SQL Server Data Tools from REMOTE

Once you receive your University of Arkansas MEC account, access will be via remote desktop connection. Remote access documentation is at the following link:

[http://enterprise.waltoncollege.uark.edu/Remote\\_Desktop\\_MEC\\_GW.pdf](http://enterprise.waltoncollege.uark.edu/Remote_Desktop_MEC_GW.pdf)

Once you're logged in to REMOTE you can use Microsoft's Business Intelligence Suite which provides tools that assist in all phases of business intelligence from building the data warehouse, creating and analyzing cubes to data mining. The following provides a data mining examples—the data mining models illustrating **classification** tasks use a table of 3333 telecommunications records. These historical records include the column, churn, which represents whether a customer left the telecommunications company or not. The idea is to build and select the best model so it can be used for predictive purposes with new customers.

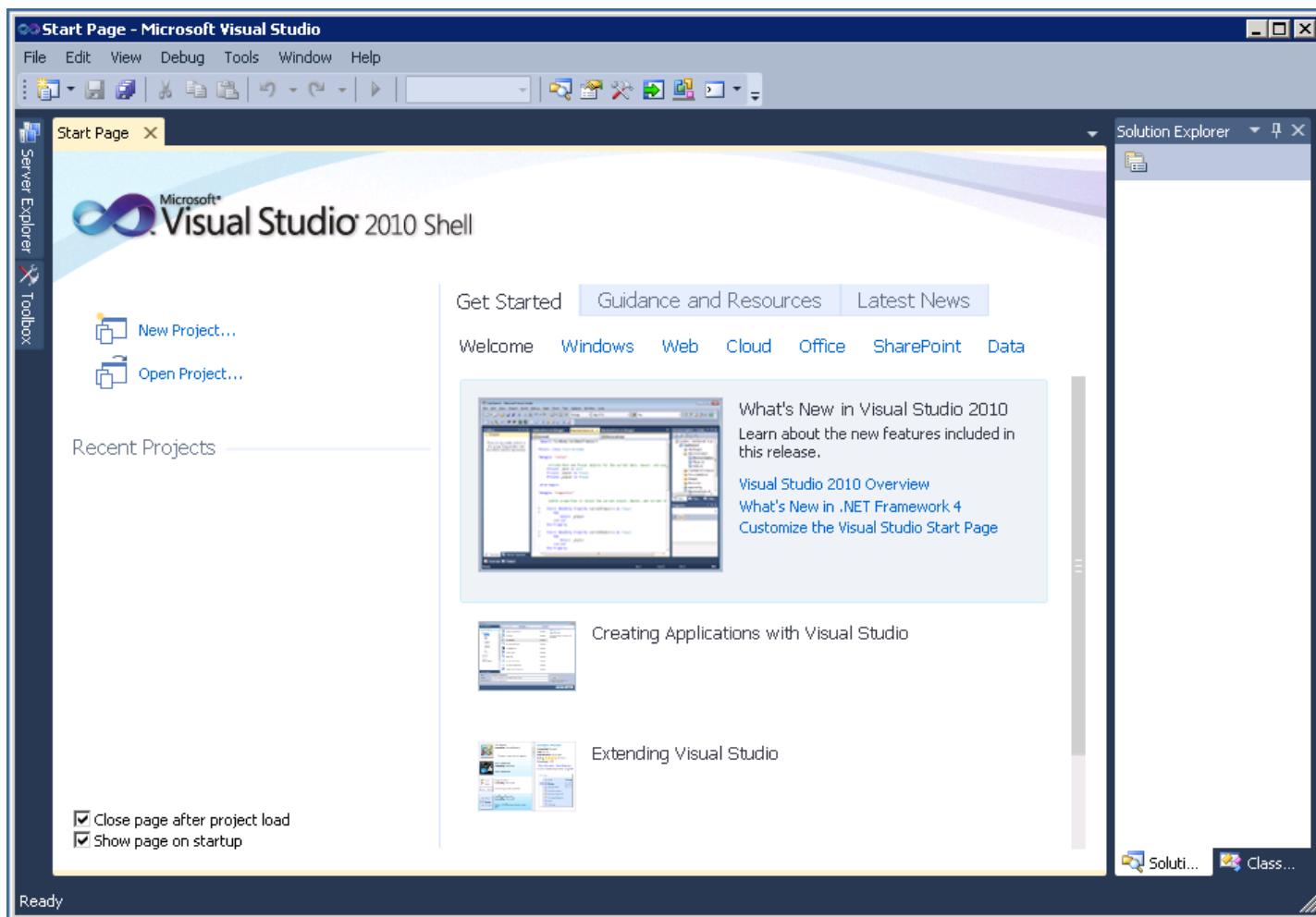
Click either the **SQL Server Data Tools** icon the Desktop or click Start and then click **SQL Server Data Tools** as shown below.



SQL Server Data Tools uses Microsoft Visual Studio (VS) as the Integrated Development Environment (IDE) which will be familiar to VB.NET or C# users. When VS opens, most likely the top will include the menu and tool bar with the Start Page tab active. Along the left of the Start page are three tabs: Getting Started, Guidance and Resources, and Latest News.

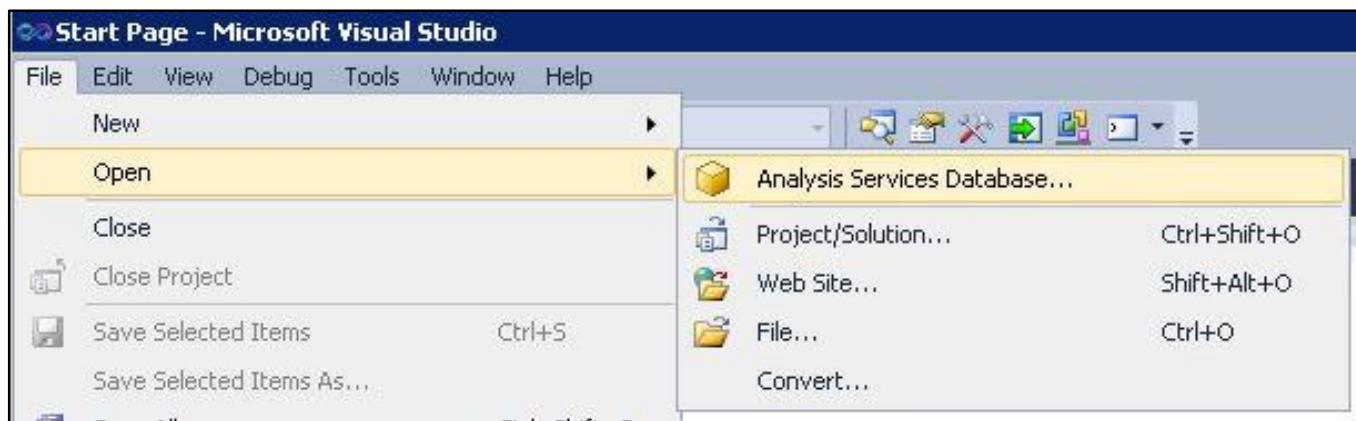
The Start button will be found at the bottom.

As usual, when you work within VS, many tabs will be created toward the top; these tabs can be closed by right-clicking and selecting Close; including the Start page tab. A partial screen shot is shown below.

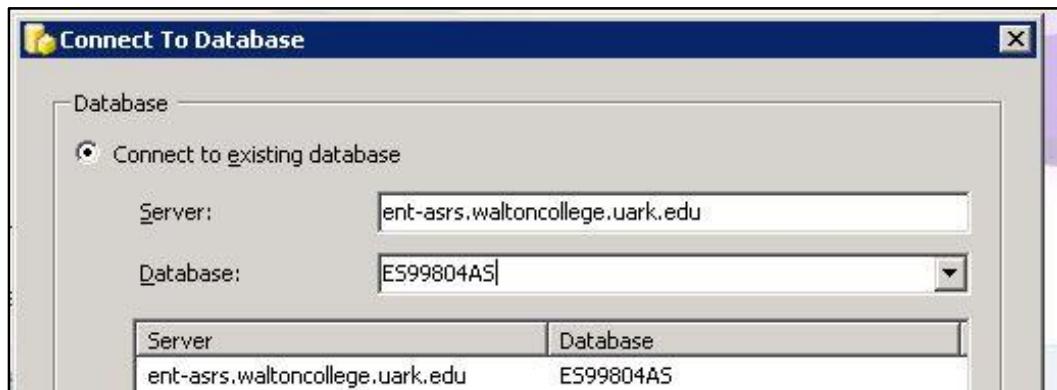


Note that you may have to scroll down using the scroll on the right to see the Start button.

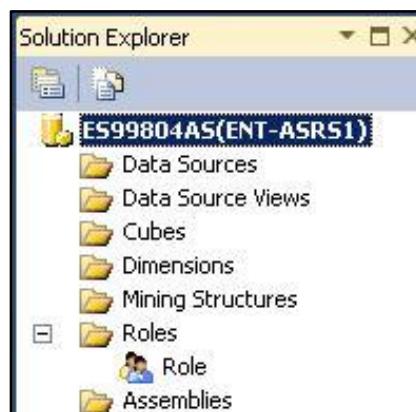
A data mining project requires using SQL Server Analysis Services—the SQL Server Analysis Server is **ENT-ASRS.waltoncollege.uark.edu**. Thus, assuming that the data to be mined is in an accessible SQL Server database (SQL Server Data Tools in this example), the first step is to connect to Analysis Services Database where you will create your BI objects. You will do this in an Analysis Services (AS) database already created for you. That AS database will have the same name as your user name with AS at the end. Example, a user with a user name ES90100 will have an AS database named ES90100AS. To connect to/access the database, click File -> Open -> Analysis Services Database...



The **Connect To Database** screen opens as shown below. Enter the Server name, **ENT-ASRS.waltoncollege.uark.edu** and press the Enter key. Use the drop down list box to select your database (account ID with AS attached at the end)—this is where Analysis Services will save your Analysis Services objects. You will only see database/s you can access (this will be your account with **AS** added with no spaces). Note that you may have to key this entry.

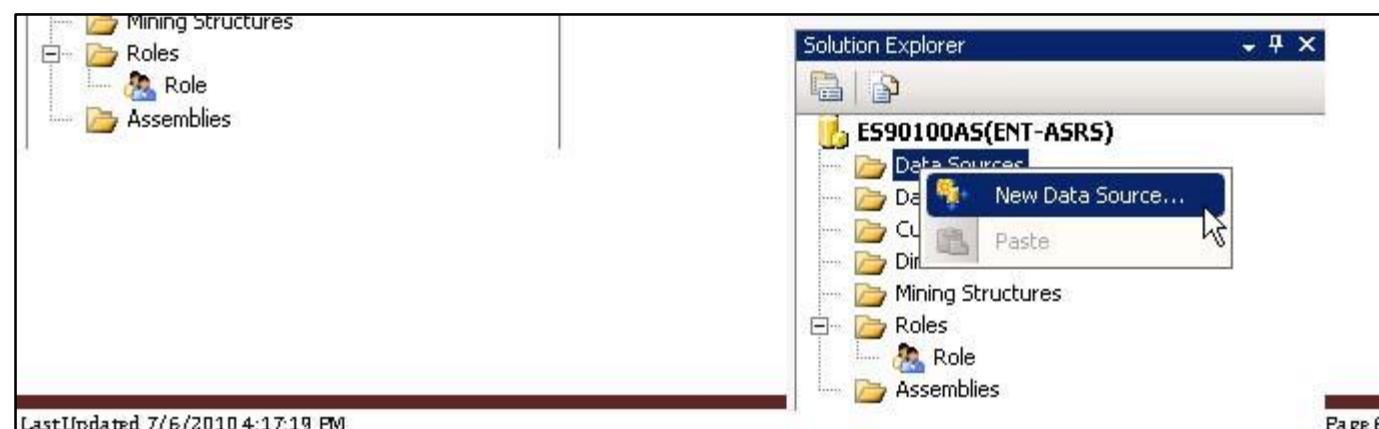


Click the OK button. Visual Studio opens – and the default location for Solution Explorer is the top right. You may need to use the horizontal scroll bar to scroll to the right to see the Solution Explorer. If it is not there, then click View on the menu and then click Solution Explorer. The name of your project should be visible with a number of other entries as shown below. The name of your project may be different from the name used in this example (doesn't matter). Your project will have the same name as the AS database you selected.

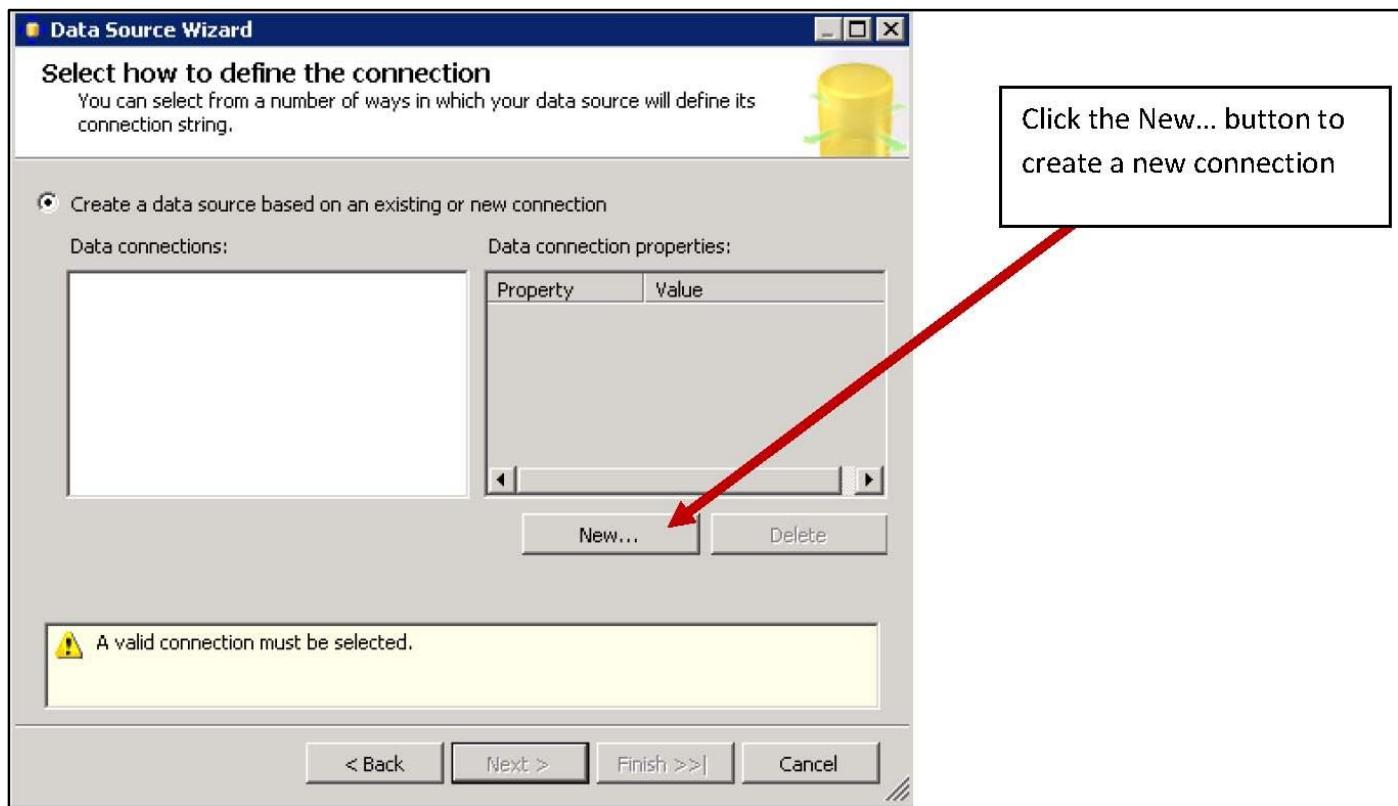


The next step requires creating a data source to be used for data mining. Thus, right-click Data Source in the Solution Explorer and click **New Data Source...**

Clicking the new Data Source option, the Data Source Wizard opens to its Welcome page. Click Next >

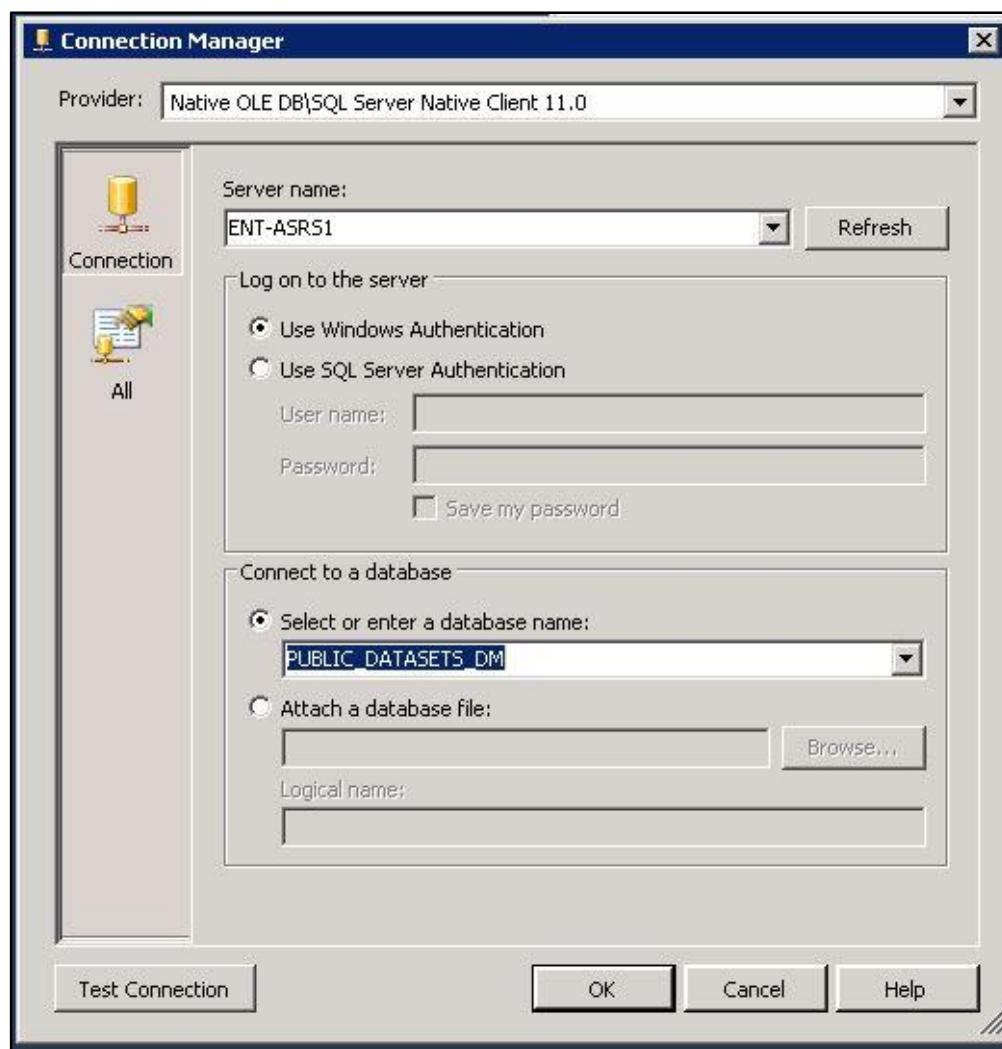


The Data Source Wizard then allows the creation of a connection by clicking the New... button.

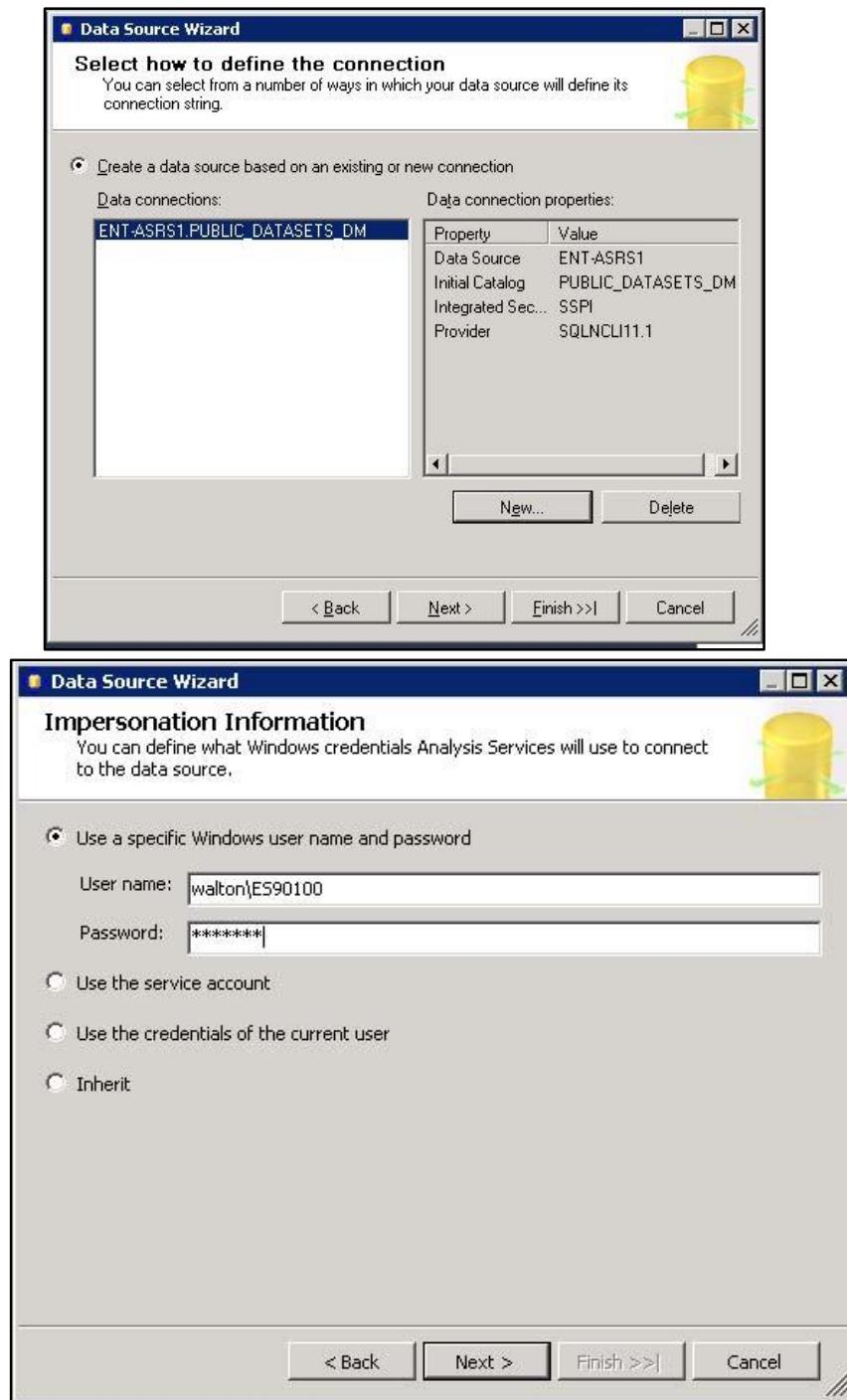


Although an existing connection may already exist, this example will create a new connection to illustrate how it works. Click the New... button. Accept the default Provider: **Native OLEDB\SQL Native Client 11.0**. Enter the Server name **ENT-ASRS1**.

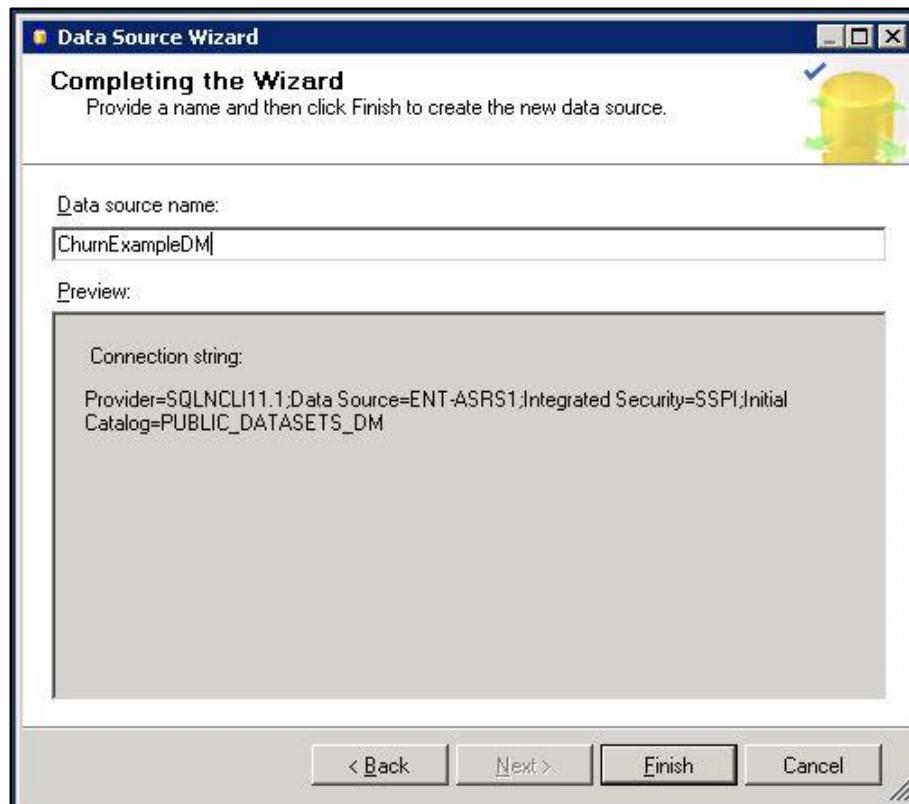
Also, use the drop down list box to select a database that has the table for data mining (for this example, the database is **PUBLIC\_DATASETS\_DM**) and click the Test Connection button (lower left) to ensure a connection exists to the database and click the OK button.



Note the Data connection properties and then click the Next button. Select **Use a specific user name and password** in the **Impersonation Information** page – enter your credentials (user name and password provided to you by the University of Arkansas)—i.e. walton\ES90100 and your password. Click the Next button.

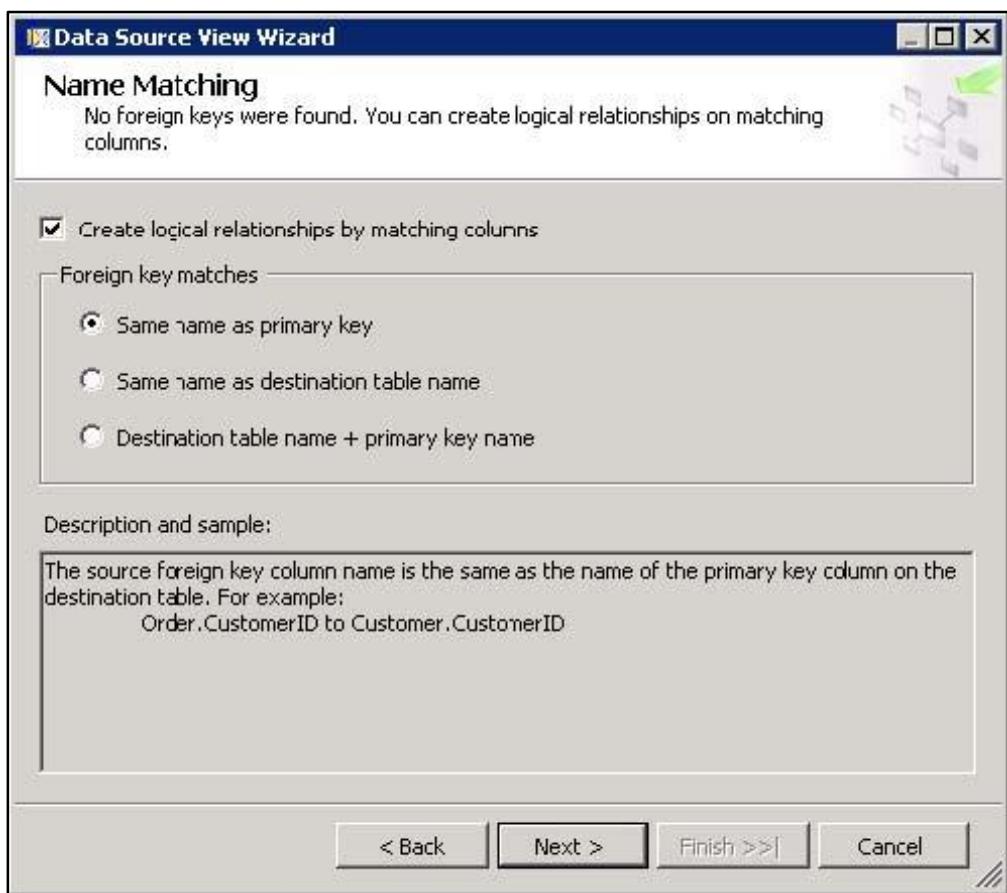


Click Finish after you provide a name to your Data Source (in this case ChurnExampleDM).



Next, a Data Source View will be needed. The Data Source View is sort of an abstract client view of the data that allows changes without affecting the original tables—a database view.

Right-click Data Source Views in the Solution Explorer and click New Data Source View to open the Data Source View Wizard. Click the Next button on the Welcome page

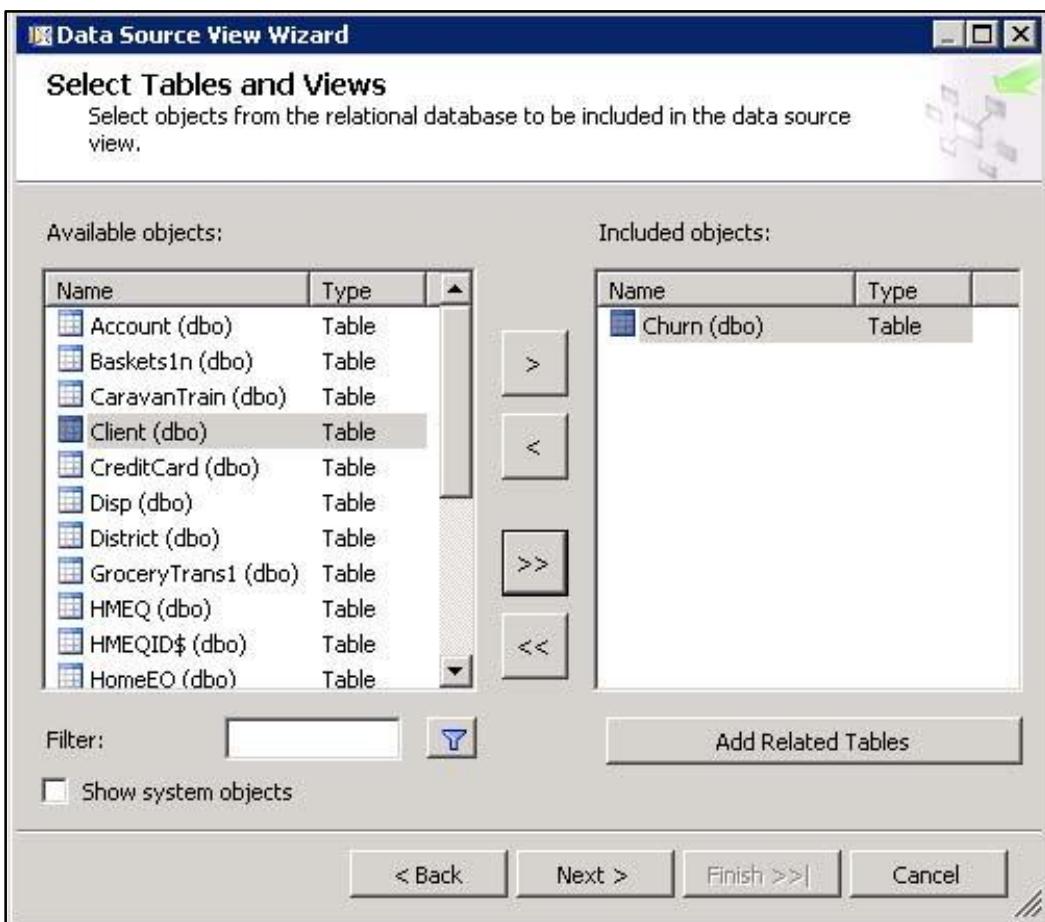


(Not shown)

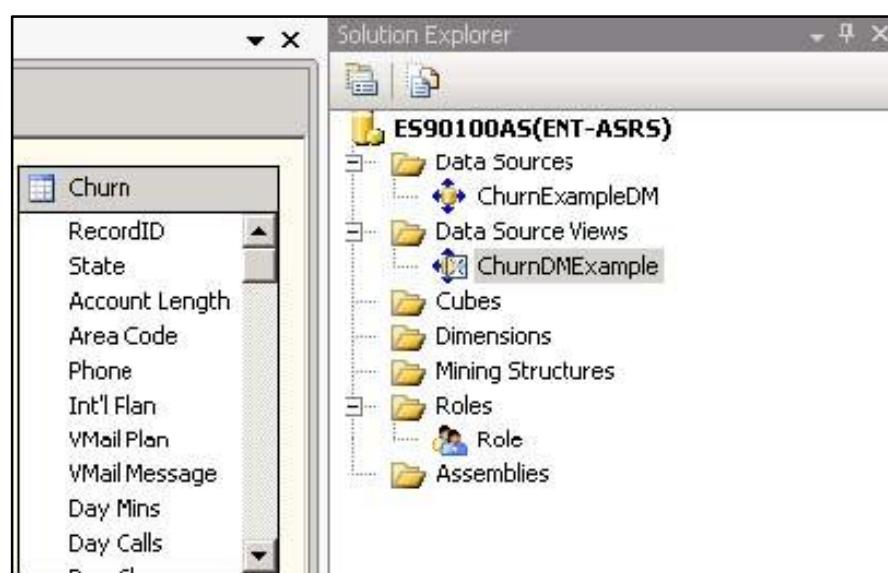
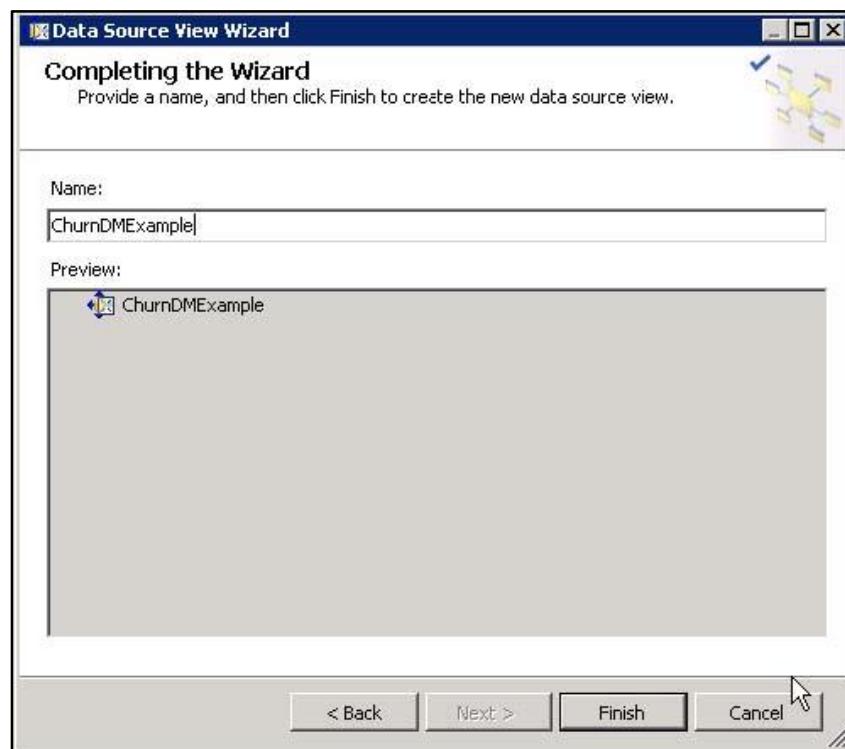
Note that the Relational data source is the one just created. Actually, this page allows creating a new data source in case one hasn't yet been created. Because the desired data source exists, click the Next button to define the Data Source View.

Ensure the **Create logical relationships by matching columns** is checked and that the foreign key matches has the **Same name as primary key** selected. Then, click the Next button.

From the **Available objects** of the **Select Tables and Views** dialog, locate and click the desired data sources in **Available objects** and click the > to move them to the list of **Included objects**. In this example, the Churn(dbo) table is the one that will be used for data mining and thus it is selected and moved to the **Included objects** list. Click the Next button.

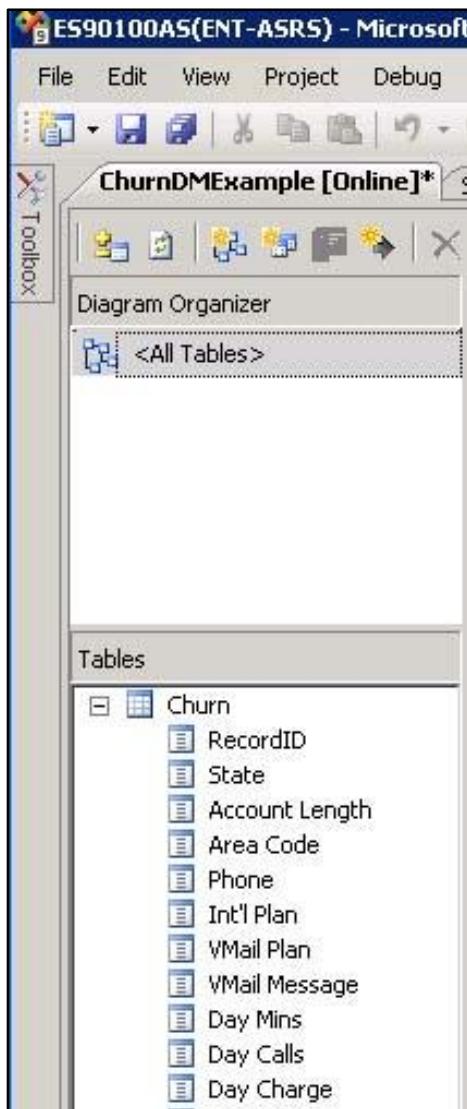


The last page of the Wizard allows you to enter a Name—enter ChurnExampleDM in the Data Source name which will be used as a data source view name in this example and click Finish.



The Data Source View is displayed as shown below. Note in the Solution Explorer, the two entries created – a data source (ChurnExampleDM) and a data source view (ChurnDMEexample) – are shown. The churn table columns are shown because the Data Source View is selected in the Solution Explorer.

On the left edge as shown on the next page, the Data Source View tab is highlighted and the Diagram Organizer and Tables are listed.



Again, this tab can be closed by right-clicking the tab and selecting Close.

**Along the way, it is always a good idea to click the Save all icon (multiple blue disks) on the tool bar. If you try to close a tab that has not been saved, it should prompt you to save your work for that part of the project.**

Now that a data source view is available, the next steps are to conduct the data mining.

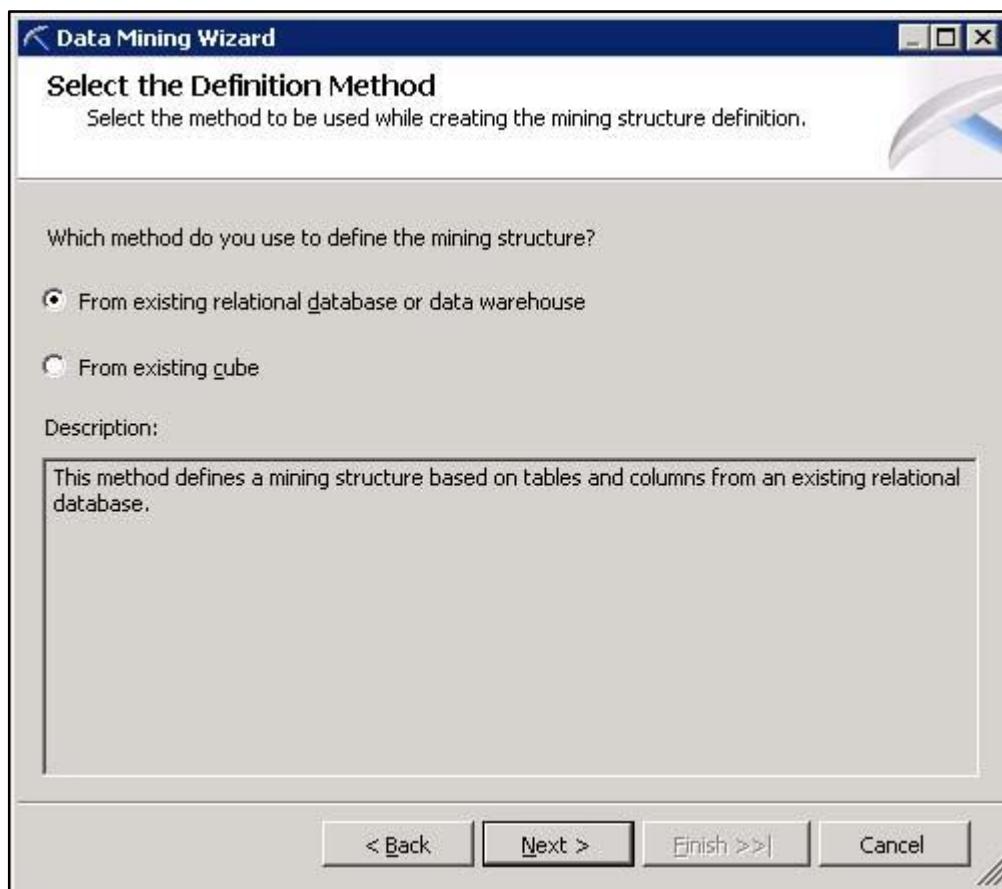
There are two parts to the data mining process – creating the mining structures and creating the mining models. The initial mining models are for classification data mining tasks and will include a **Decision Tree**, **Logistic Regression**, and **Neural Network** model. These models will then be compared to determine the best model. The first model will be a decision tree.

The data mining structure defines the domain of a data mining problem and the data mining model involves the algorithm to run against the data. First create a mining structure by right-clicking Mining Structures in the Solution Explorer window and selecting Create New Mining Structure which opens the Data Mining Wizard.

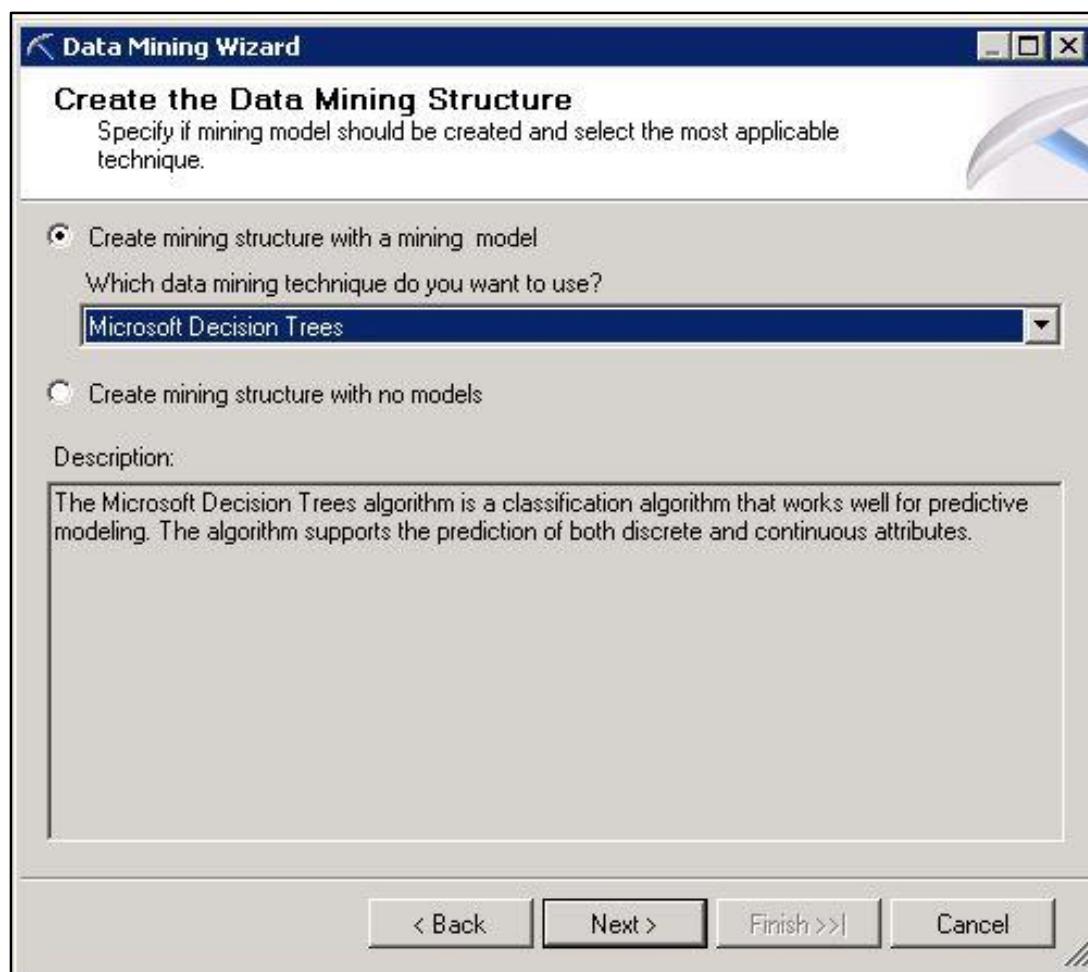
Create a new mining structure by right-clicking Mining Structure in Solution Explorer which opens the **Data Mining Wizard**. Click the Next button on the Welcome page (not shown) to get to the Select the Definition Method.



Accept the default Definition Method option: **From existing relational database or data warehouse** and click the Next button.



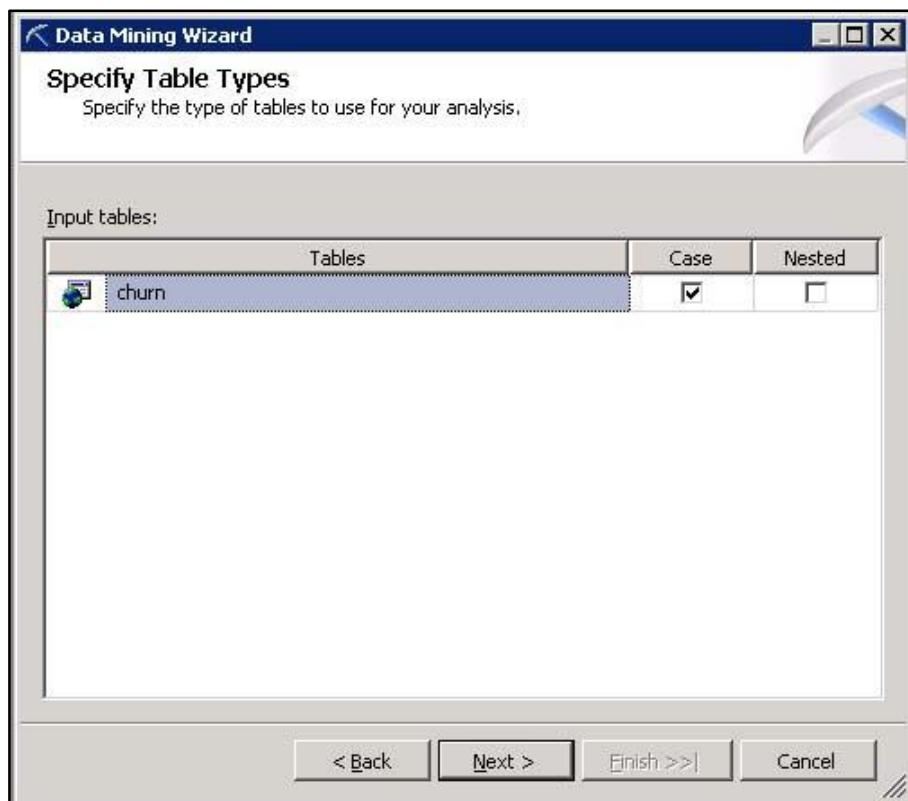
The default data mining technique is Microsoft Decision Trees—note, use the drop down list box to select other data mining techniques. Click the Next button because this example will use a decision tree analysis.



The Select Data Source View (not shown) page already displays the most recently created Data Source View. Note that if other Data Source Views have been created, they can be located via the Browse button. Because the desired Data Source View is selected, click the Next button.

The **Specify Table Types** page defaults to the churn table as defined in the Data Source View. Also, the format of the churn table is Case—each record represents one customer record. The Nested format allows directly using multiple tables in a relational database.

For this example, Case is the correct format so click the Next button.



The next page is the **Specify the Training Data** page. This allows specifying the columns to be used in the analysis and also the target variable for supervised (directed) data mining tasks – **Classification** in this illustration. Decision trees, logistic regression, and neural networks (classification) are directed data mining tasks and the target variable will be the variable Churn?. The Churn? column has true or false as its values representing whether the customer left the company (true) or not (false).

The **Specify the Training Data** page lists each column name of the churn table on a row which allows it to be specified as a **Key**, **Input**, or **Predictable**. Note that both the Case format and the Nested format require a column to be specified as a Key. For this example, the RecordID will be specified as the key.

Also, note that in this data mining example, the purpose is to be able to predict those that will churn or not based on column input values. Thus, the variable Churn? is selected as a predictable variable.

Not all the columns in the churn table will be used in the data mining analysis. From exploratory data analysis (not shown), it was determined that the variables (columns) State, Area Code and Phone contained bad data. Also, all the columns related to Charge were perfectly correlated to the corresponding Mins (Minutes) column so none of the Charge columns will be used in the analysis.

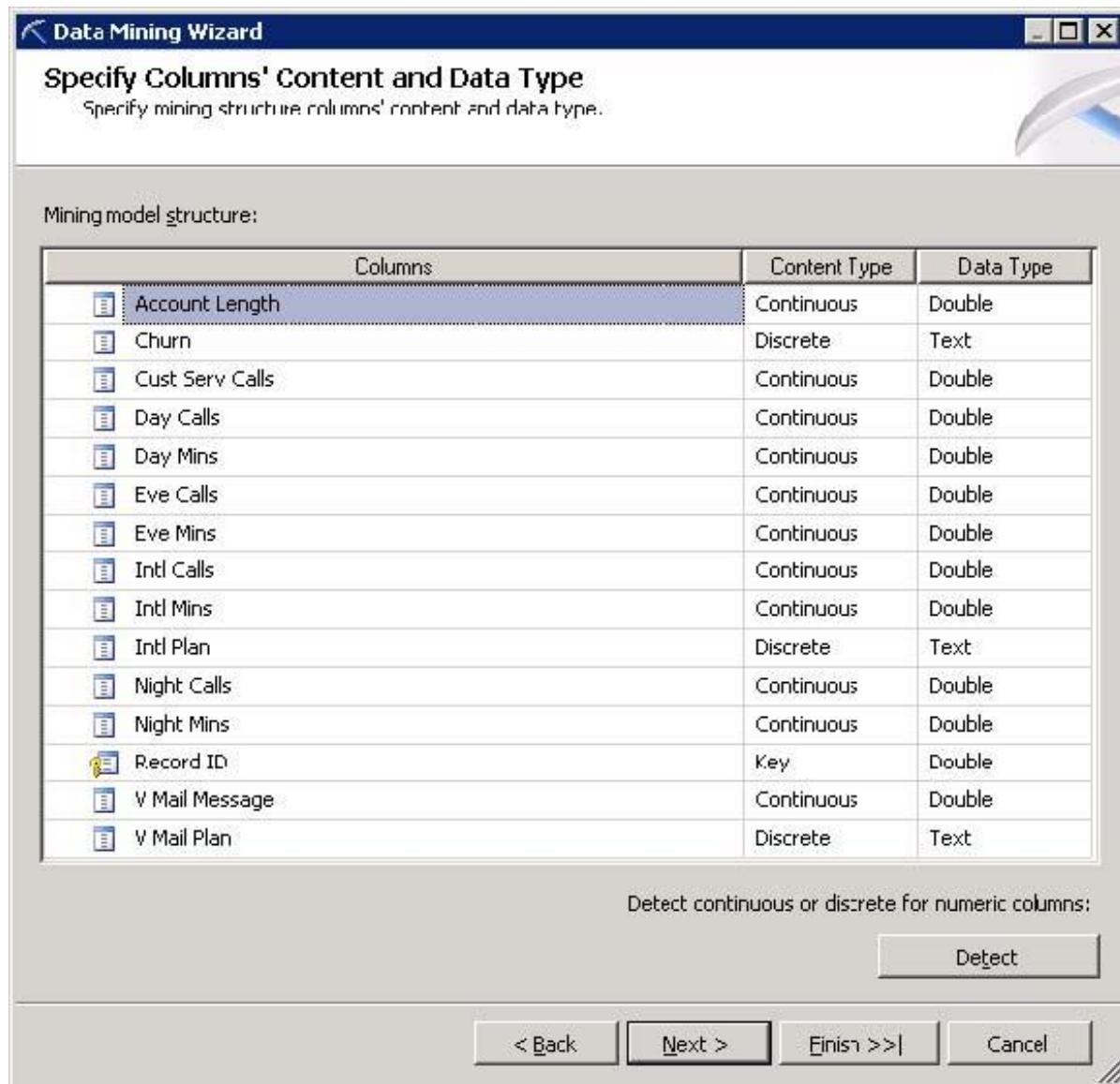
The screenshot shows the 'Data Mining Wizard' window with the title 'Specify the Training Data'. The sub-section is 'Mining model structure'. A table lists columns from the 'Churn' table, categorized by icon: Account Length, Area Code, Churn?, CustServ Calls, Day Calls, Day Charge, Day Mins, Eve Calls, Eve Charge, Eve Mins, Intl Calls, Intl Charge, Intl Mins, Int'l Plan, Night Calls, Night Charge, Night Mins, Phone, RecordID, State, VMail Message, and VMail Plan. Columns are checked for 'Input' and 'Predict...' status. The 'Churn?' column is checked for both. The 'RecordID' column is checked for 'Key'. The 'Suggest' button is visible at the bottom right.

	Tables/Columns	Key	Input	Predict...
Churn				
Account Length	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Area Code	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Churn?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
CustServ Calls	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Day Calls	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Day Charge	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Day Mins	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Eve Calls	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Eve Charge	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Eve Mins	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Intl Calls	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Intl Charge	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Intl Mins	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Int'l Plan	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Night Calls	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Night Charge	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Night Mins	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Phone	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
RecordID	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
State	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
VMail Message	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
VMail Plan	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>

Recommend inputs for currently selected predictable:  
Suggest

Notice the **Suggest** button—lower right—will suggest which variables should be included as input variables.

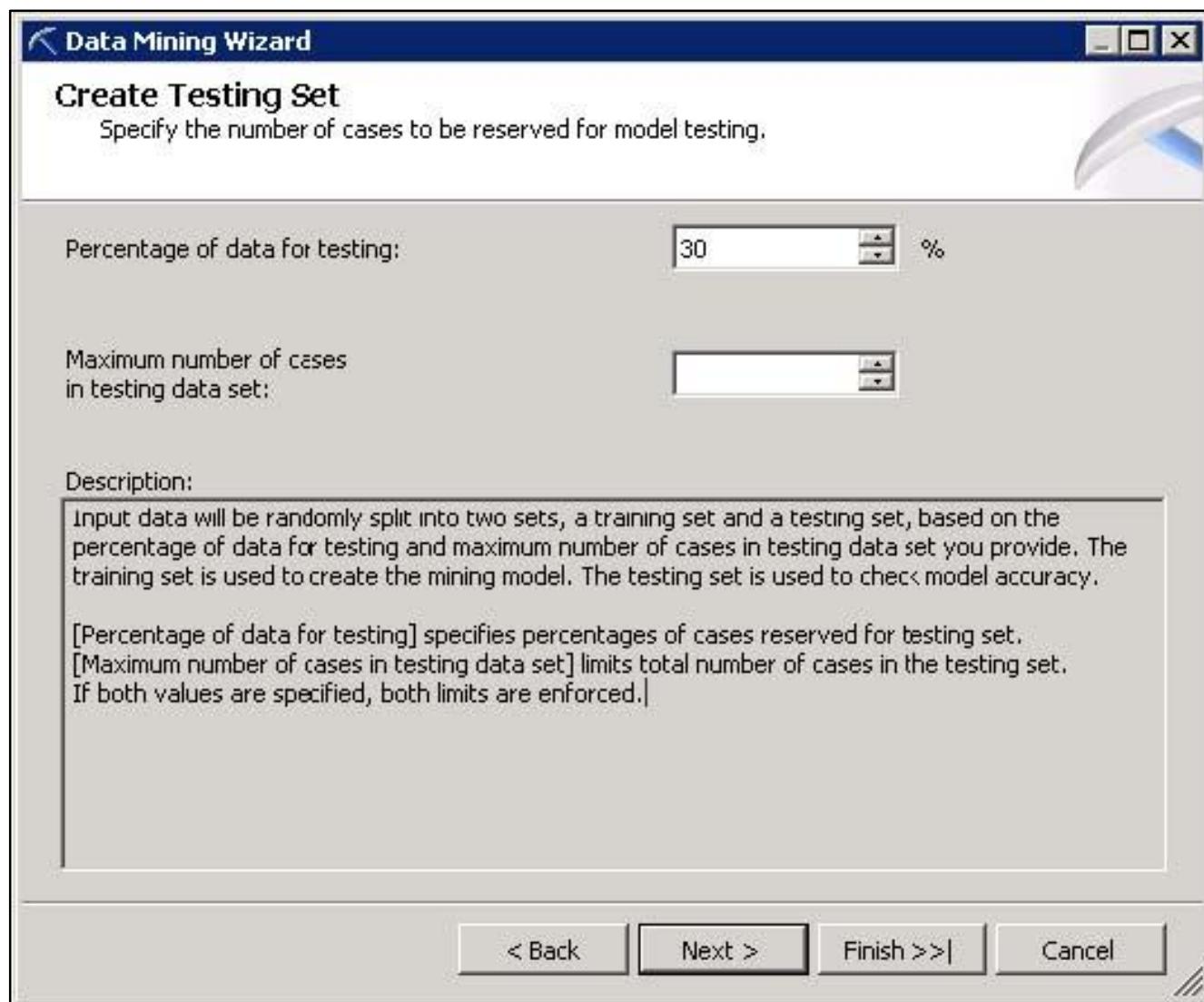
Click the Next button which displays a page indicating the data types of the variables to be used in the data mining algorithm.



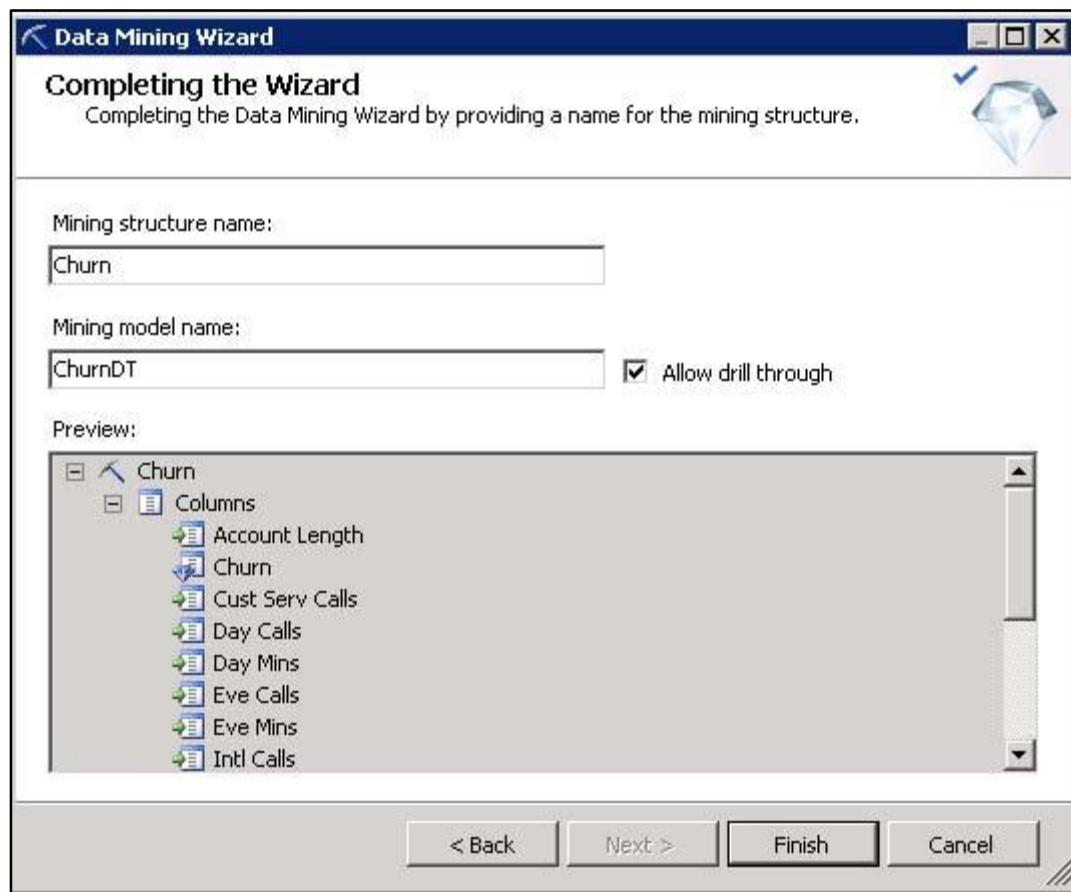
Notice that the Churn? variable is discrete and needs to be as the objective of the data mining algorithm is to have a model (decision tree in this case) that will predict Churn? as true or false.

Click the Next button which allows partitioning the data into a training set and a test set. Having a test set for a model built on a training set is very important. It provides information on the stability of the model and also the generalization of the model.

Accept the default 30% random test value – resulting in a training set of 70% of the records.



Click the Next button to the last page of the Data Mining Wizard—Completing the Wizard. The user can provide a name for the mining structure and a name for the mining model. In this example, Churn is used for the name of the mining structure and ChurnDT is used for this particular Decision Tree model. Click the Finish button.



It is important to Process the project. From the Solution Explorer, right-click the Mining Structure entry and click Process.

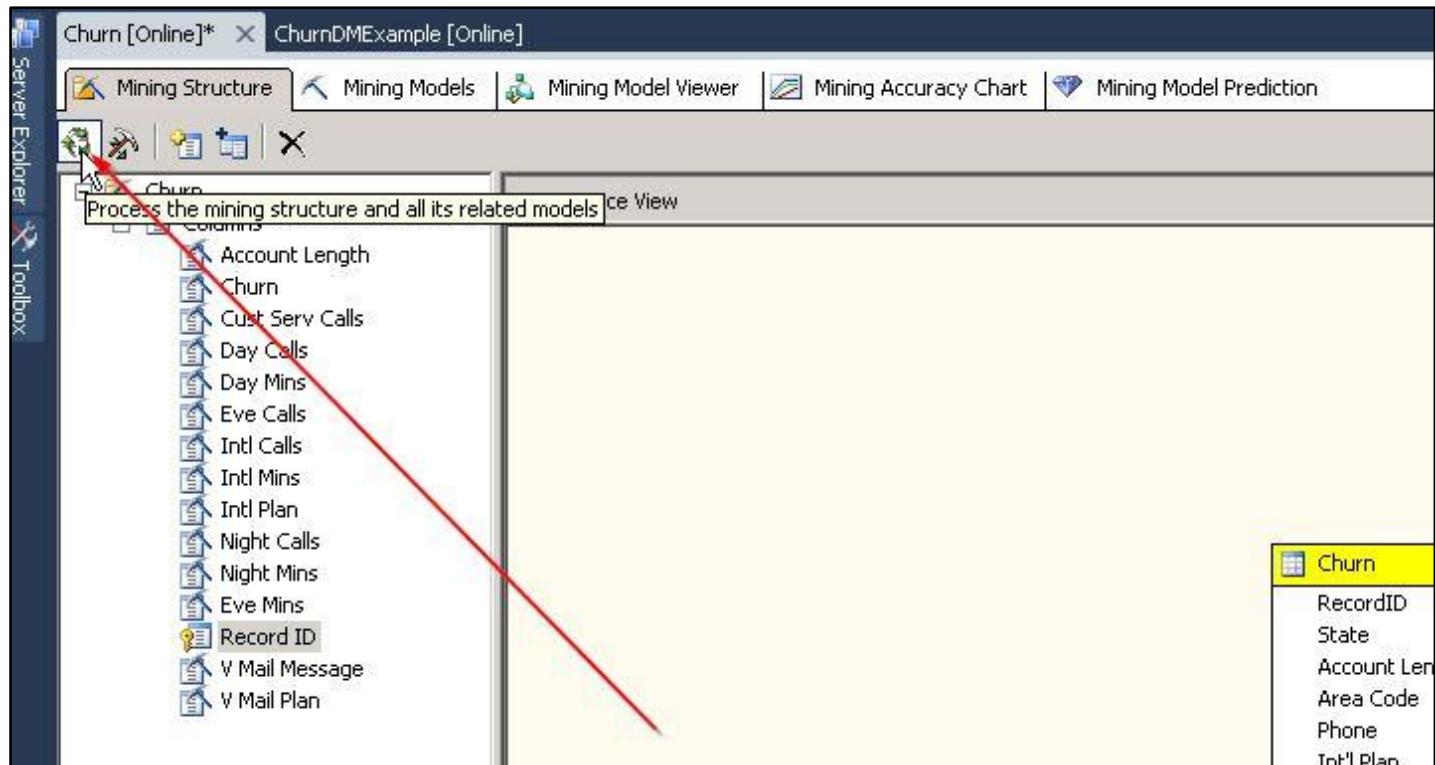


If you have not saved since making changes; you will be prompted to save all changes before processing. Click the Yes button. Click the Run... button (not shown) on the Process Mining Structure-Churn dialog. Processing may take a bit of time so be patient. The system confirms that Process Succeeded or lists errors if there is a problem. Click the Close button after the Process completes and then Close to exit the Process Mining Structure dialog.

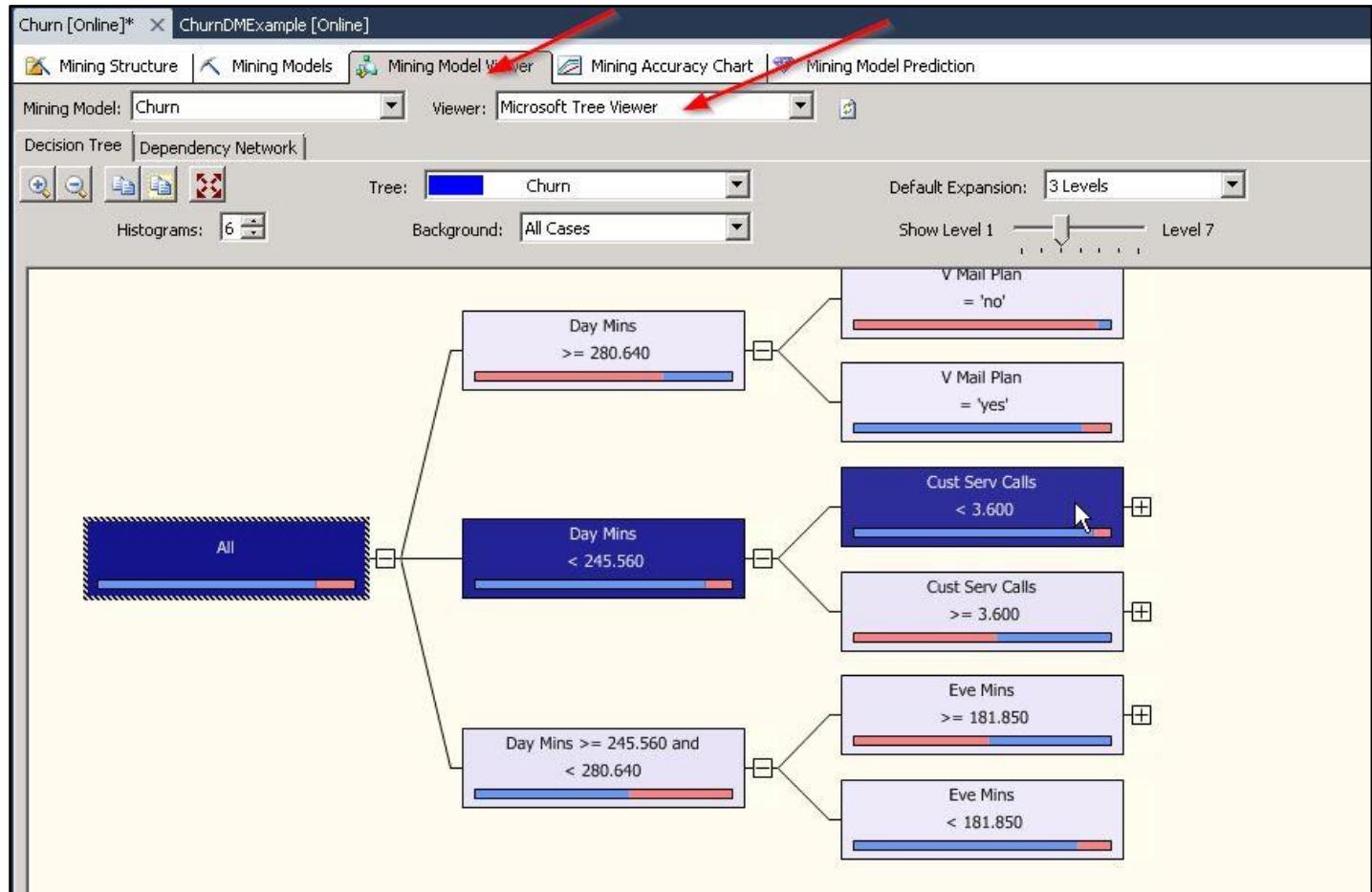
Review the top left of Visual Studio. Again, many tabs may be present—the ChurnDMEExample.dmm[Design], ChurnDMEExample.dsv[Design], Start Page and the row underneath includes the Mining Structure, Mining Models, Mining Model Viewer, Mining Accuracy Chart, Mining Model Prediction.

Clicking the Mining Models tab provides a summary of the model—recall this is Decision Tree model named ChurnDMEExample.

Notice the green circular icon directly above the structure column. The icons on this row allow adding and processing data mining models—this green circular icon processes one or more data mining models. If this decision tree model has not been run, then click this icon; the mining structure may request being run again.



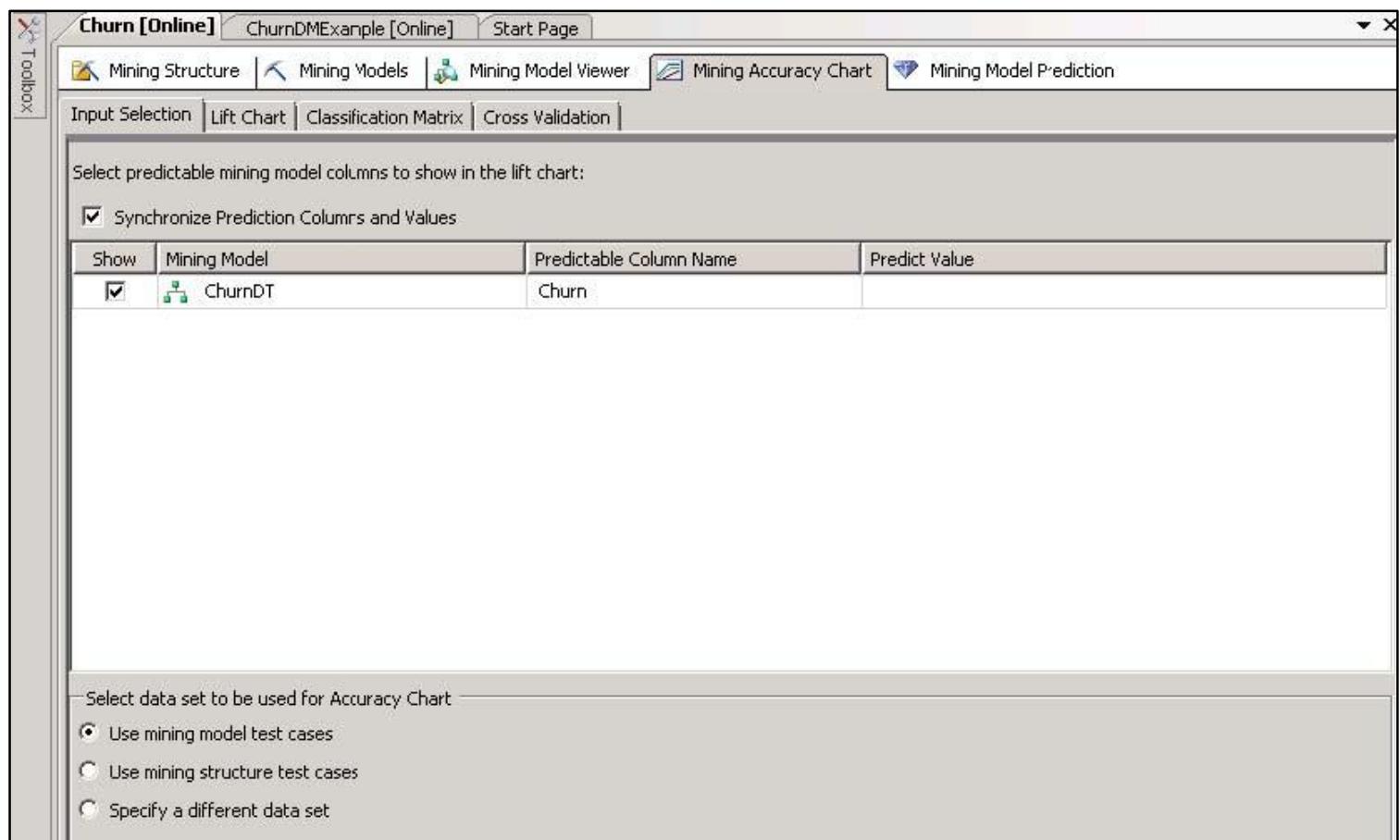
After the model has run, click the **Mining Model Viewer** tab and **Microsoft Tree Viewer** from the **Viewer** drop down list box.



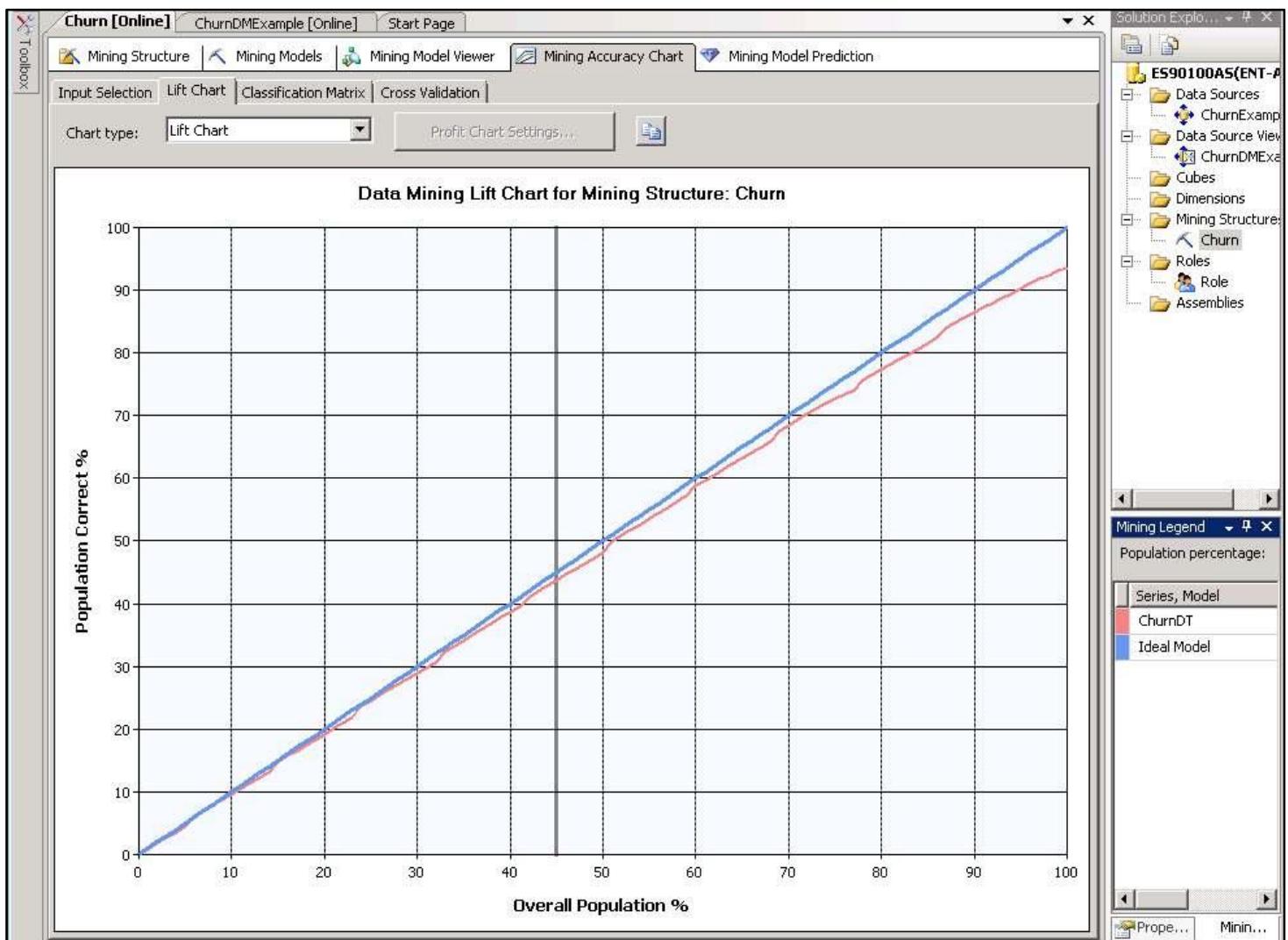
The tree viewer provides options for viewing the tree—for example the default number of levels to display. Also, the **Background** drop down list box allows different refinements – in this case, the possibilities are False, Missing, and True as well as the default for all cases.

Notice that moving the mouse over the bar on a tree node provides data about that node. The one illustrated here is a leaf node—no other nodes follow it—and it has 56 cases of which 53 are true and 3 are false. Thus, those that have Day Mins > than 280.64 and also have a Voice Mail Plan are highly likely to churn.

Click the Mining Accuracy Chart tab.



Click the Lift Chart tab and select Lift Chart from the Chart type drop down list box.



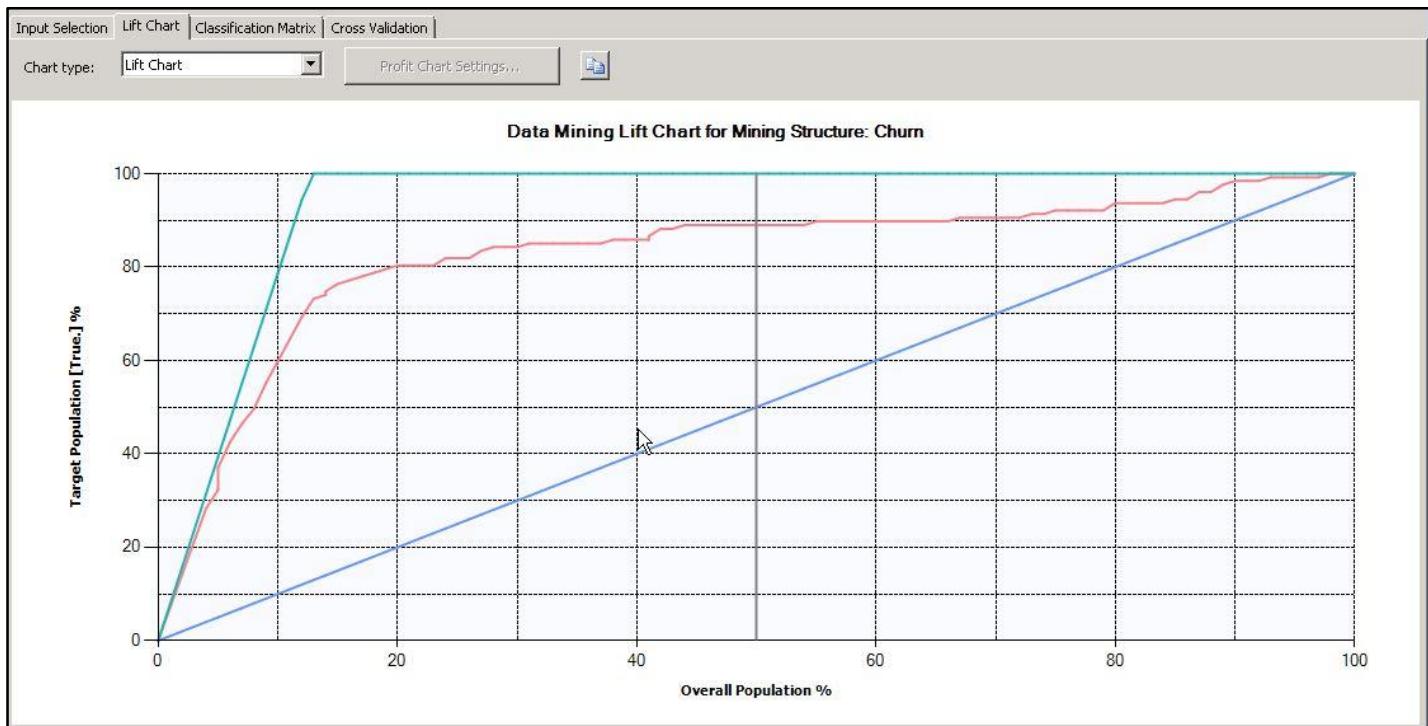
The blue diagonal line represents an ideal model and the red line shows the results of the decision tree model—the Score is the accuracy of the model based on the training dataset which is 97.4%—see lower right hand window named Mining Legend.

One can also obtain a more traditional lift chart by populating the **Predict Value column in the Input Selection tab**—this part of the window is shown below.

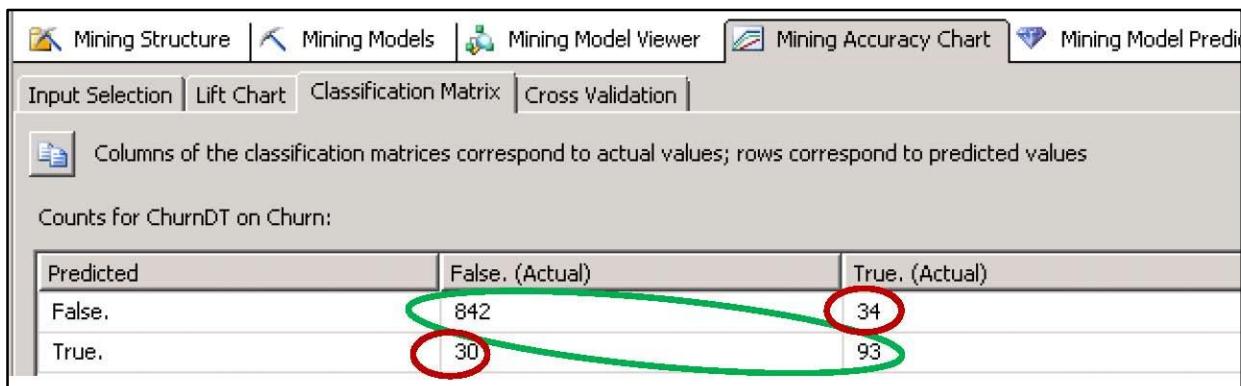
This screenshot shows the 'Input Selection' tab of the Data Mining ribbon. It displays a table for selecting predictable mining model columns. The table has columns: 'Show', 'Mining Model', 'Predictable Column Name', and 'Predict Value'. There is one row selected, showing 'ChurnDt' in the 'Mining Model' column, 'Churn' in the 'Predictable Column Name' column, and 'True.' in the 'Predict Value' column. A checkbox 'Synchronize Prediction Columns and Values' is checked. A note above the table says: 'Select predictable mining model columns to show in the lift chart:'.

Show	Mining Model	Predictable Column Name	Predict Value
<input checked="" type="checkbox"/>	ChurnDt	Churn	True.

Click the **Lift Chart** tab to get a lift chart based on a Predict Value of True as shown below. The green line is an ideal model, the red line is the decision tree and the blue line would be the result with random guess. The decision tree model tracks fairly well the ideal model for this training set.



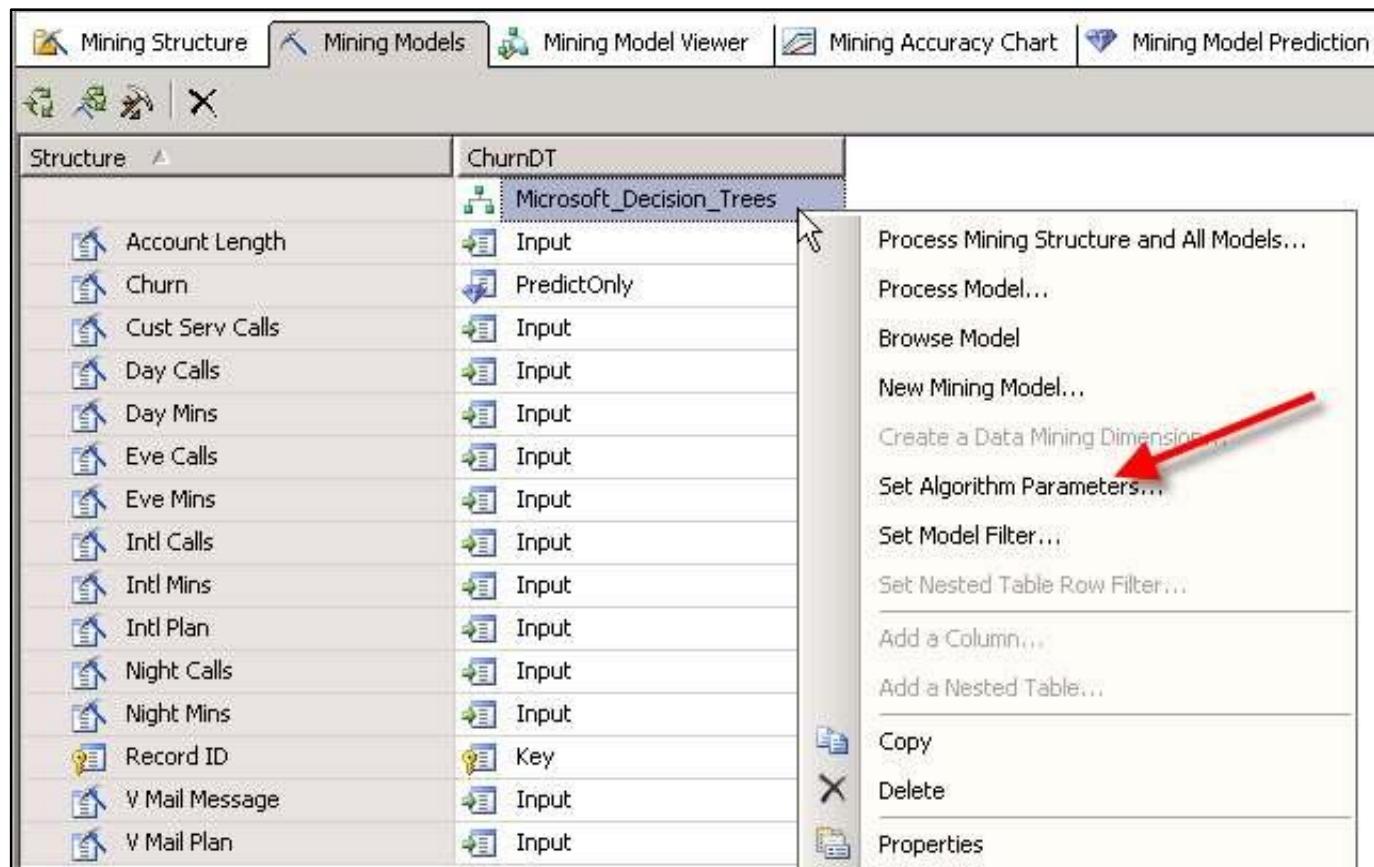
Click the **Classification Matrix** sub-tab to view a table of the model's predicted values versus actual values. This model predicted 842 cases as False that were in fact False but also predicted 34 as False that were actually True; the model predicted 93 cases as True when in fact they were True but also predicted 30 cases as True when in fact they were False. Thus, the diagonal values circled in green represent where the model correctly predicted the actual values and they should be considerably larger than the off-diagonal values which represent where the model missed predicting the actual values.

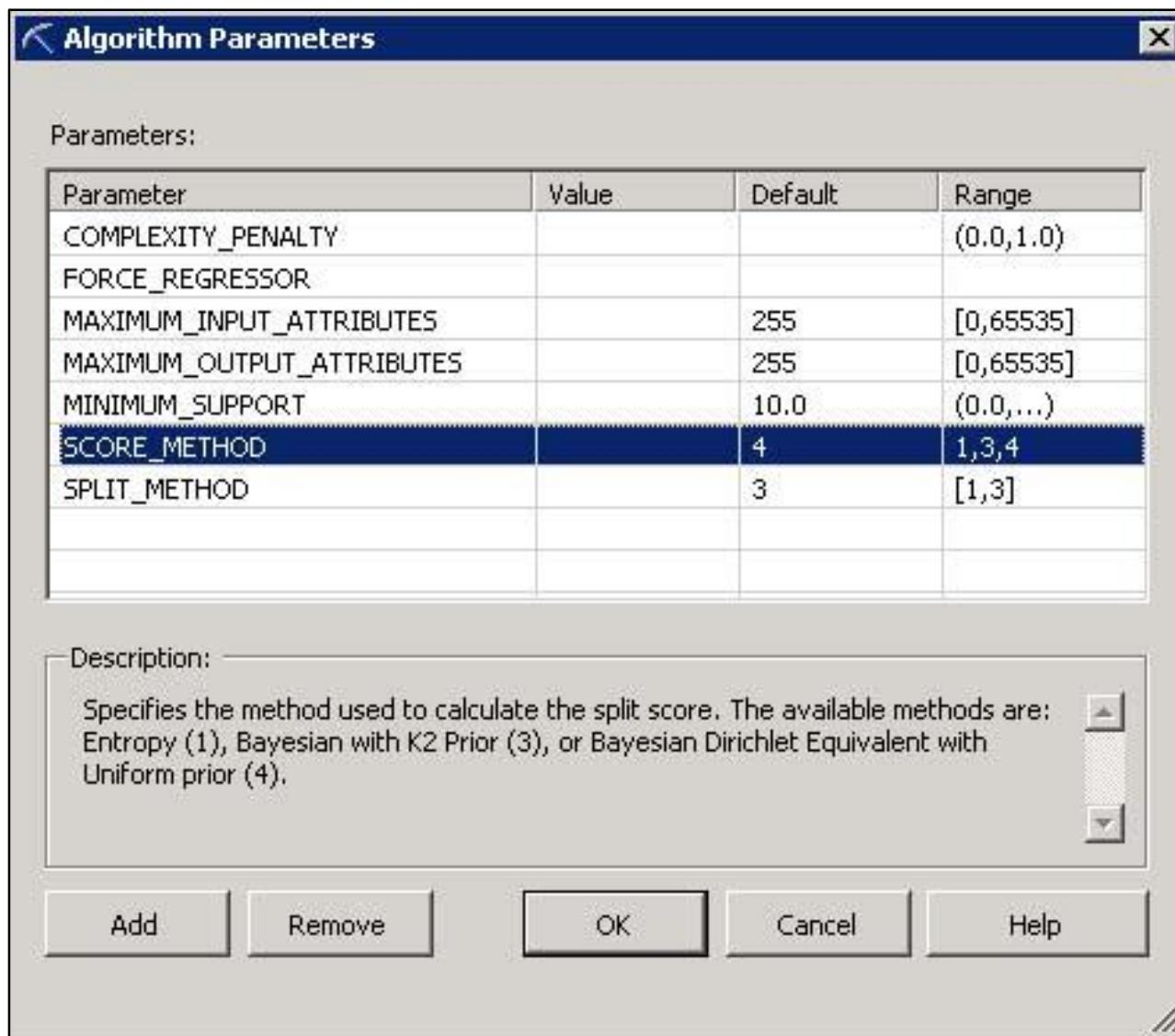


The values circled in red are values the model predicted incorrectly. The 34 value is referred to as a False Positive—meaning that the model predicted 170 cases as False when in fact they were True. The 30 value is referred to as a False Negative as the model predicted it to be True when in fact it was False. Note that the impact of a false positive and a false negative may be greatly different. Microsoft allows you to take these differences into account via a cost matrix which assigns cost values to the outcomes.

## Mining Model Parameters

The intent of this tutorial is not to teach the mining algorithms; rather, it is to provide examples of using Microsoft's Business Intelligence Development Studio for data mining. Data mining algorithms generally have parameters that can be set by the user to improve model development. The user can change selected default parameters of data mining algorithms by right-clicking the data mining model and selecting **Set Algorithm Parameters**.

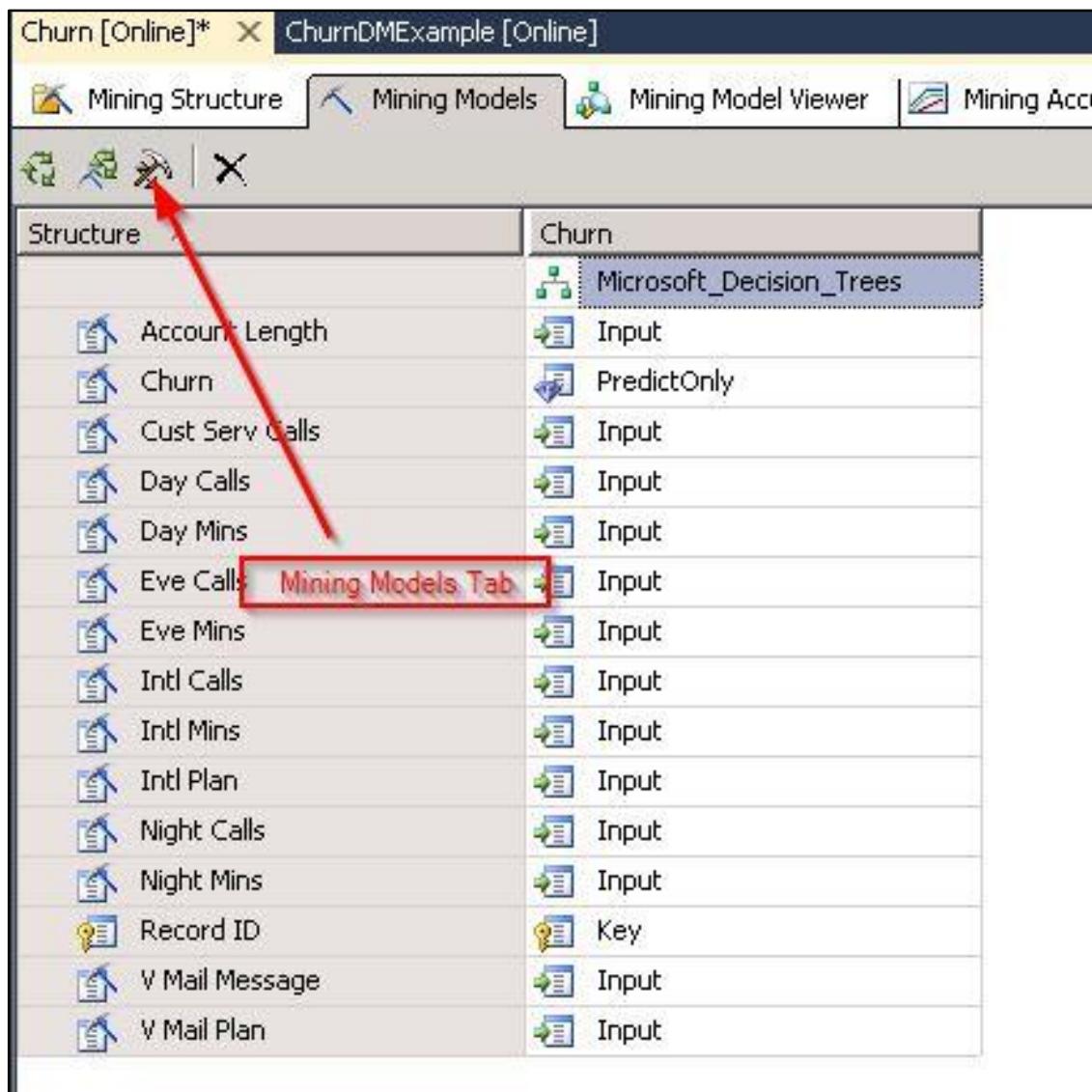




The Algorithm Parameters Window above shows the parameters the user can set for decision trees. The particular row shown indicates a Bayesian Dirichlet Equivalent with Uniform prior method is the default SCORE\_METHOD. Change the default setting to 1 to use the Entropy method.

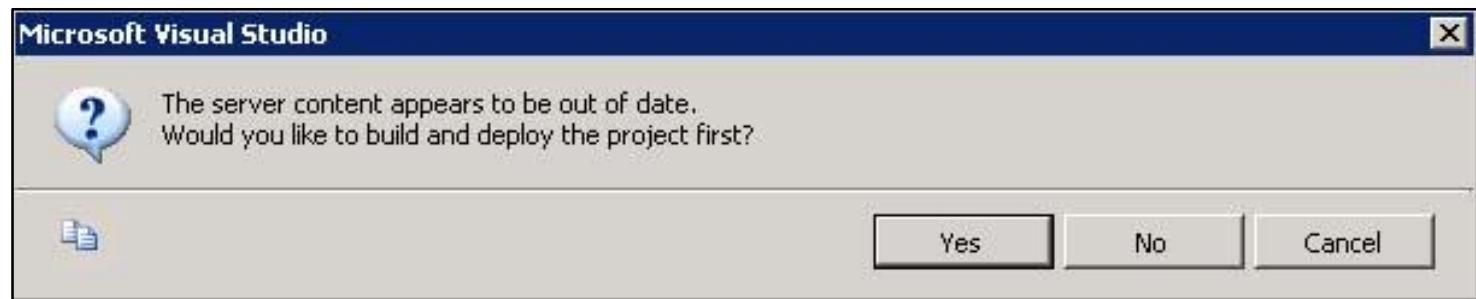
## Adding Additional Supervised (Directed) Classification Models

Additional data mining algorithms for this classification task can be easily added and compared. For example, neural networks and logistic regression also are used for creating classification models. To add a model, click the Mining Models tab and then click the hammer/chisel icon (new model) to open a dialog which allows you to provide a name and the data mining algorithm you wish to run.



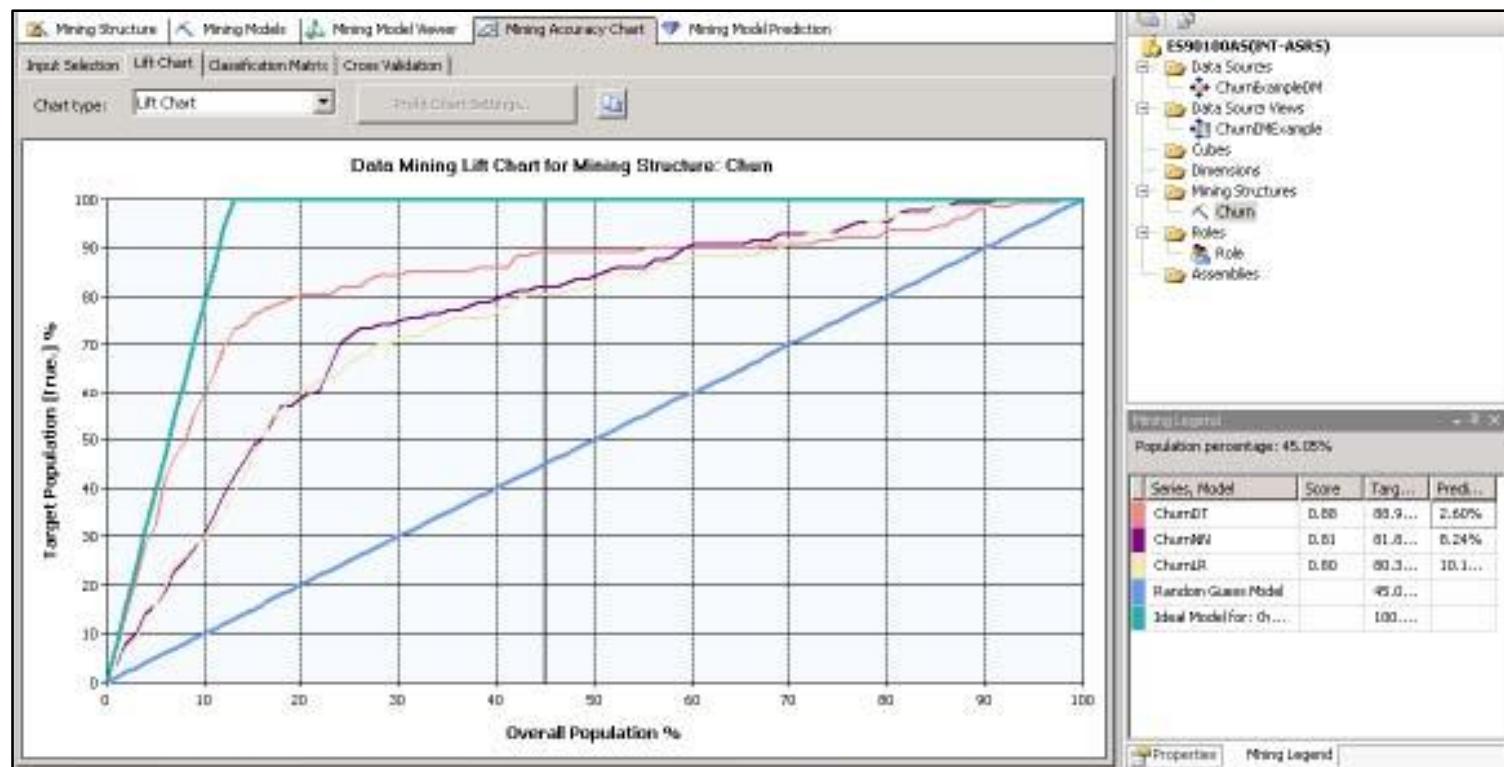
In this example, a neural network and a logistic regression model have been added. Click the green circle icon to run all the models.

You will get a prompt that indicates the models are out of date, click Yes to build and deploy the project with the new models.



Then click the Run button on the next dialog. Note that you may need to click Yes at times to keep the model up to date and deployed.

The easiest way to compare the classification models is by lift and percent accuracy. Click the Mining Accuracy tab and note the lift chart for each model compared to an ideal model—shown below is for prediction of True. Also, in the lower right-hand pane, each model has a score that represents the model's accuracy—in this case the decision tree model is superior in terms of model accuracy for this data.



Recall that the more traditional lift chart shown above was viewed based on setting a predict value. As shown below, all three models have the Predict Value set to True—enforced via checking the **Synchronize Prediction Columns and Values**.

Select predictable mining model columns to show in the lift chart:

Synchronize Prediction Columns and Values

Show	Mining Model	Predictable Column Name	Predict Value
<input checked="" type="checkbox"/>	ChurnDt	Churn	True,
<input checked="" type="checkbox"/>	ChurnNN	Churn	True,
<input checked="" type="checkbox"/>	ChurnLogReg	Churn	True,

The classification matrices for the three mining models are shown below.

Churn [Online] ChurnDMExample [Online] Start Page

Mining Structure Mining Models Mining Model Viewer Mining Accuracy Chart Mining Model Prediction

Input Selection Lift Chart Classification Matrix Cross Validation

Columns of the classification matrices correspond to actual values; rows correspond to predicted values

Counts for ChurnDT on Churn:

Predicted	False. (Actual)	True. (Actual)
False.	842	34
True.	30	93

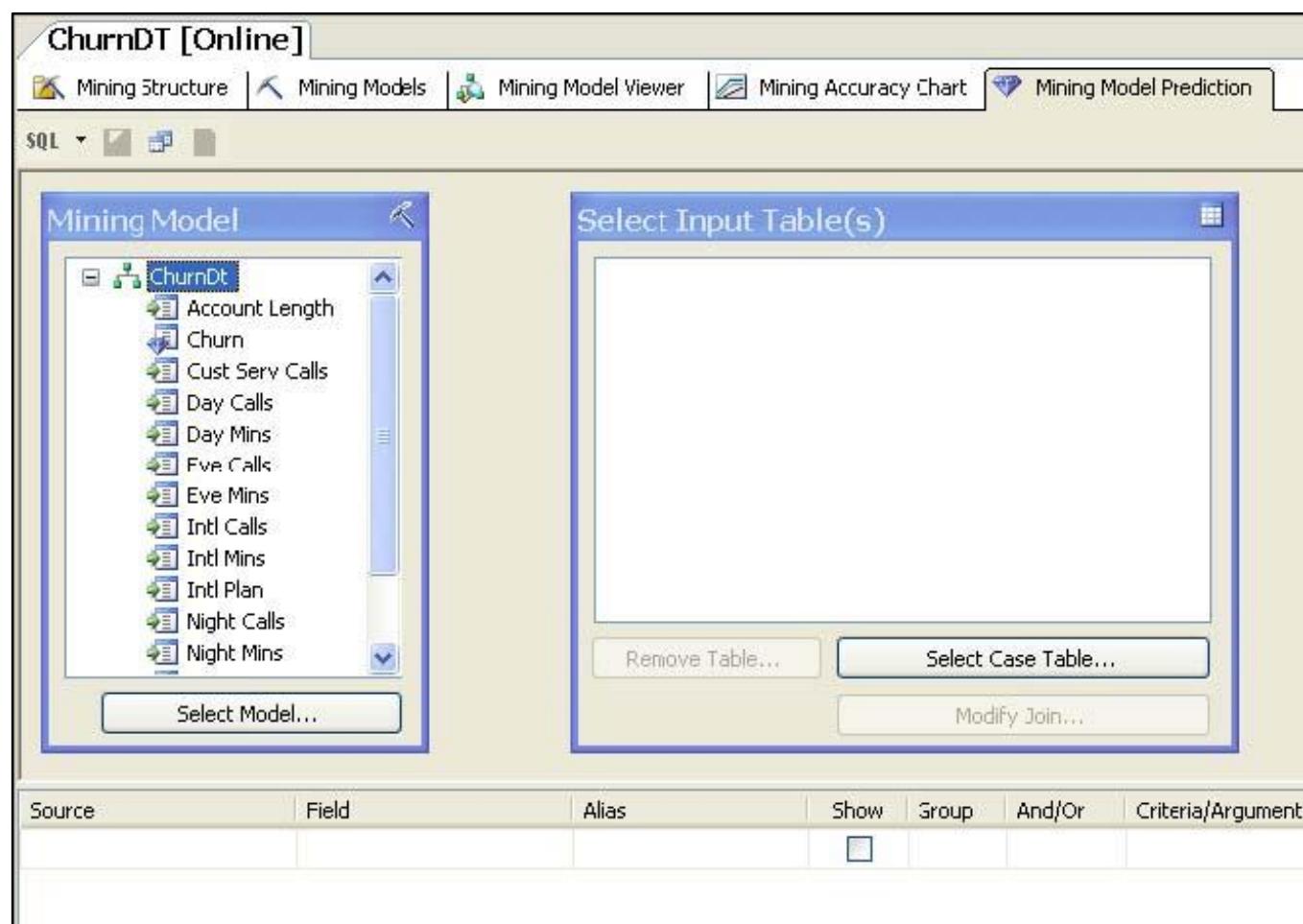
Counts for ChurnNN on Churn:

Predicted	False. (Actual)	True. (Actual)
False.	847	108
True.	25	19

Counts for ChurnLR on Churn:

Predicted	False. (Actual)	True. (Actual)
False.	845	106
True.	27	21

Although not illustrated here, the usual resulting decisions of building these classification models is to select the best performing model—may be based on cost values instead of just misclassification rate—and apply it new data. Clicking the **Mining Model Prediction** tab opens the window shown below. This window allows the user to select a model and then the data that the model will be applied to.



## Undirected (unsupervised) Data Mining

Association Analysis, typically used for Market Basket Analysis (MBA), and Clustering are two data mining tasks that fall into this category of data mining. Clustering will be illustrated using the same churn data as was used with the classification models (decision tree, neural network, and logistic regression). Because MBA has received so much attention in the data mining literature and practice press, an example using purchase data will be used to illustrate association analysis.

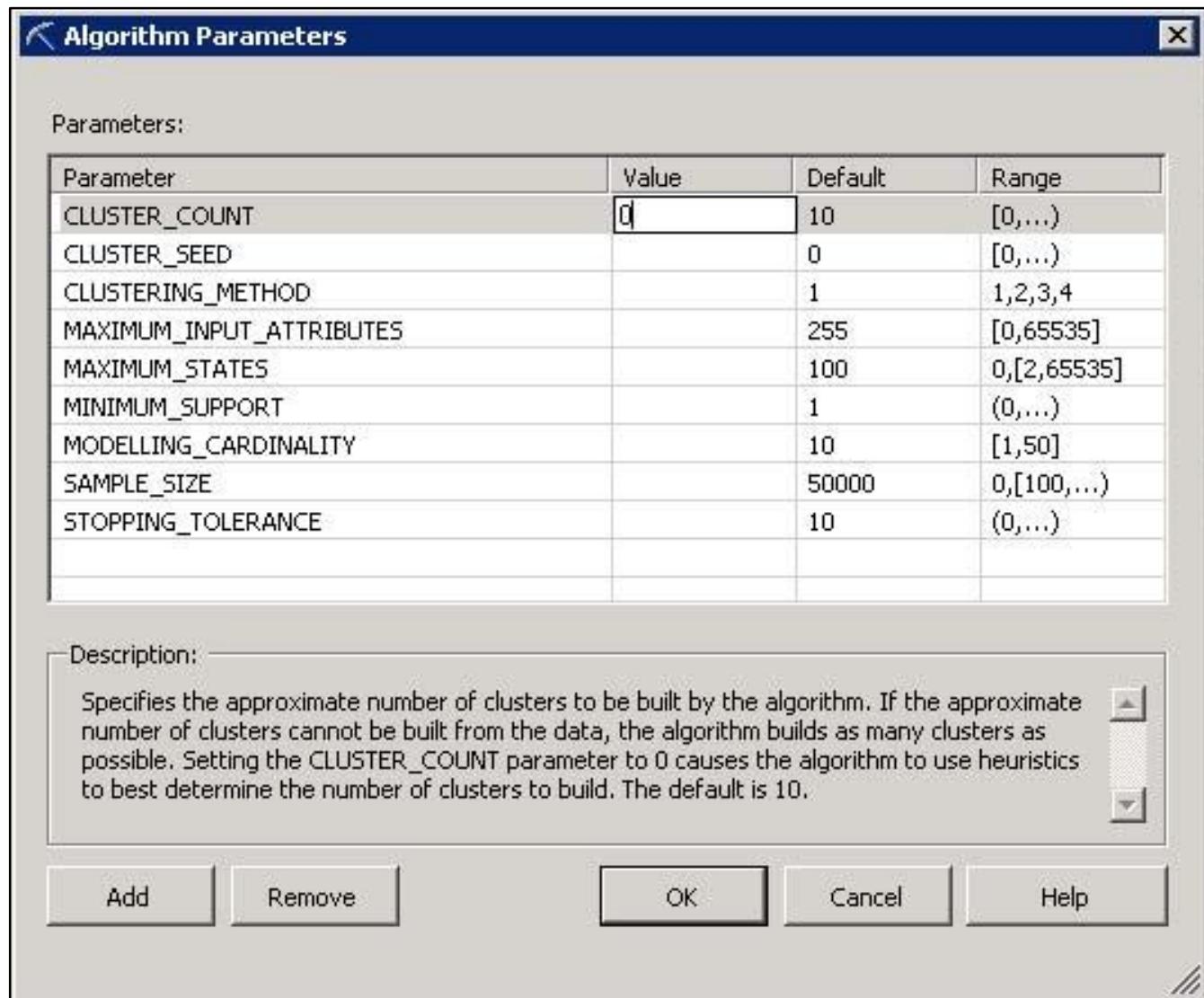
Clustering, typically a non-supervised data mining task does not have a target variable. The churn data has an obvious target variable churn, but two approaches using clustering may be helpful in the data mining process. Recall that clustering is the process of grouping records (cases) into similar clusters based on their attributes. One approach would be to leave the churn variable in for clustering—this approach may lead to insights about the attributes of the cases and the variable churn. Another approach is to leave out the churn variable for clustering and then add the cluster number to each record (case) that can be used downstream for classification tasks. Either way, clustering typically is not an end (there are exceptions); rather an exploratory process. Insights learned in the clustering process can be used for further model development.

There is not an easy way to determine how many clusters to build. Thus, experimenting with the number of clusters may be necessary for exploring the data. Finally, interpretation and understanding of clusters generally requires significant domain knowledge.

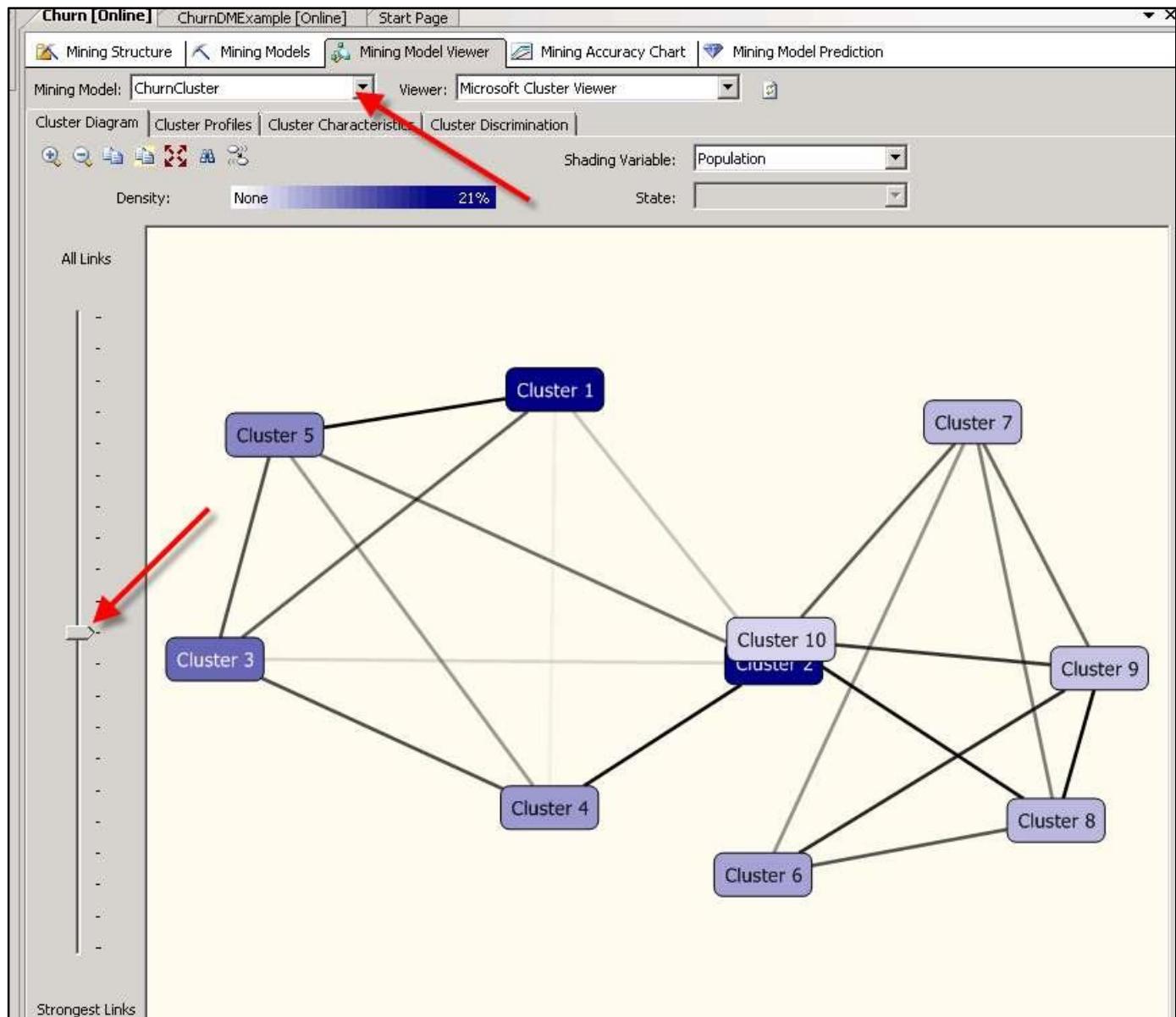
To illustrate clustering, add a new mining model as shown below—note that the churn variable is included as PredictOnly. Clustering can be run by leaving Churn? in the analysis to see what cluster(s) it is associated. Also, one may wish to remove Churn?, run the clustering algorithm, add a cluster number to each record and then run a classification model. It may make the classification model stronger.

Structure	Decision_Trees	ChurnNN	ChurnLR	ChurnCluster
Account Length		Microsoft_Neural_Network	Microsoft_Logistic_Regression	Microsoft_Clustering
Churn		PredictOnly	PredictOnly	PredictOnly
Cust Serv Calls		Input	Input	Input
Day Calls		Input	Input	Input
Day Mins		Input	Input	Input
Eve Calls		Input	Input	Input
Eve Mins		Input	Input	Input
Intl Calls		Input	Input	Input
Intl Mins		Input	Input	Input
Intl Plan		Input	Input	Input
Night Calls		Input	Input	Input
Night Mins		Input	Input	Input
Record ID		Key	Key	Key
V Mail Message		Input	Input	Input
V Mail Plan		Input	Input	Input

Right click the cluster model, select **Set Algorithm Parameters** and set the CLUSTER\_COUNT value to zero in the Value column—heuristics will be used to help determine the number of clusters—the default value is 10.



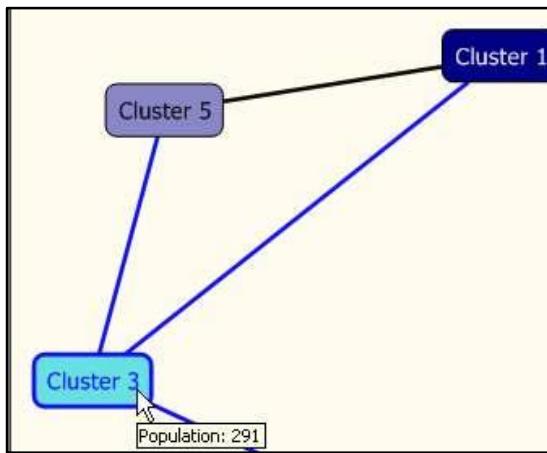
Run the model and view the results—major tab is Mining Model Viewer and the model selected for viewing is ChurnCluster from the dropdown menu. There are four sub tabs—Cluster Diagram, Cluster Profiles, Cluster Characteristics and Cluster Discrimination. The default tab is Cluster diagram. \*This may generate an error if the project is not saved.



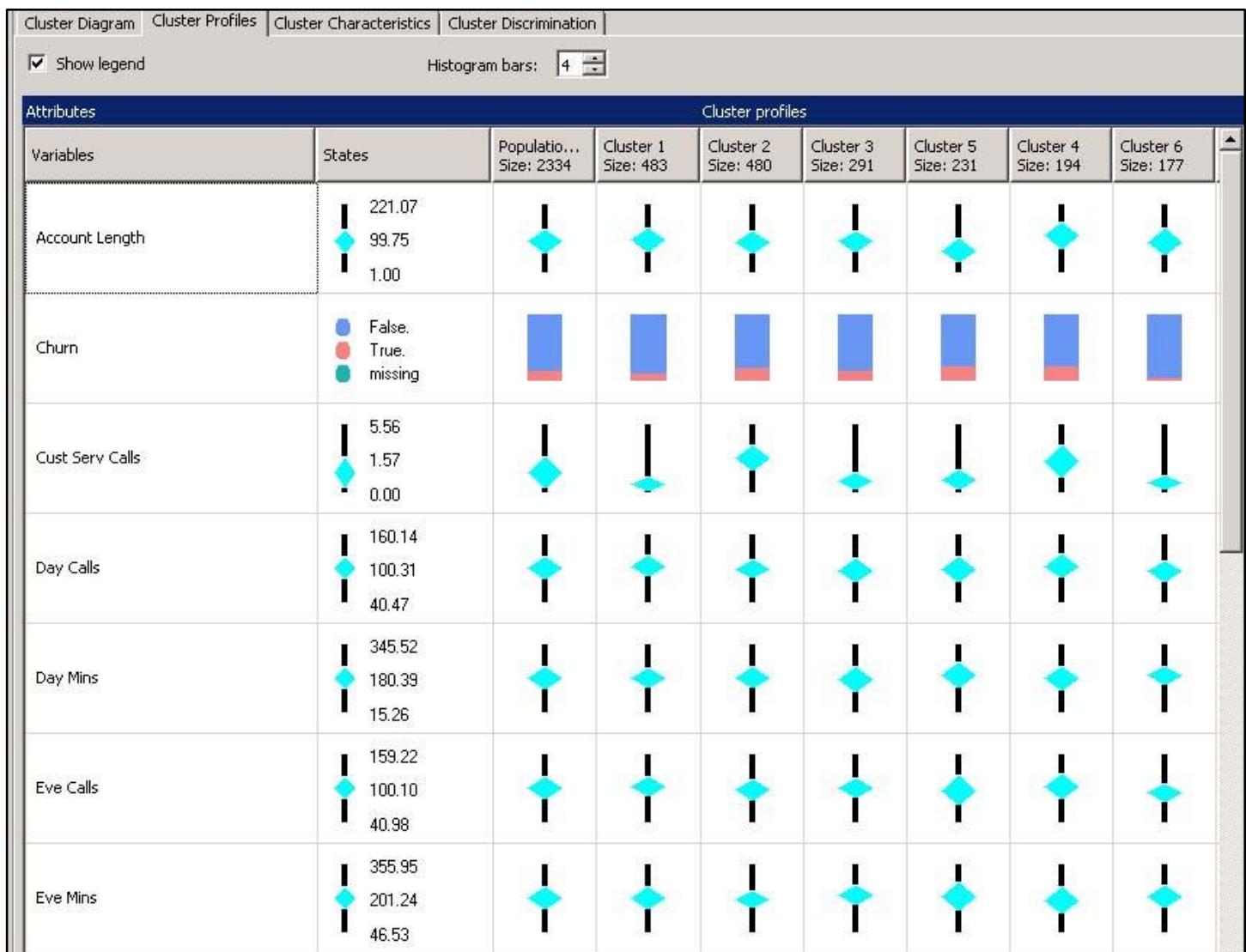
The slider on the left is set about half way between including all links between clusters and only the strongest links. Move the slider to the top and then to the bottom to see the links change—line density is a measure of strength. Also, moving the mouse cursor over a cluster indicated how many records (cases) are in the cluster.

Note that 10 clusters have been produced and that the default names for the clusters are cluster 1, cluster 2, cluster 3... The challenge is to review the records (cases) in each cluster to determine which clusters, if any, may have important new and usable information.

Clicking Cluster 3 produces the following and contains 414 records.



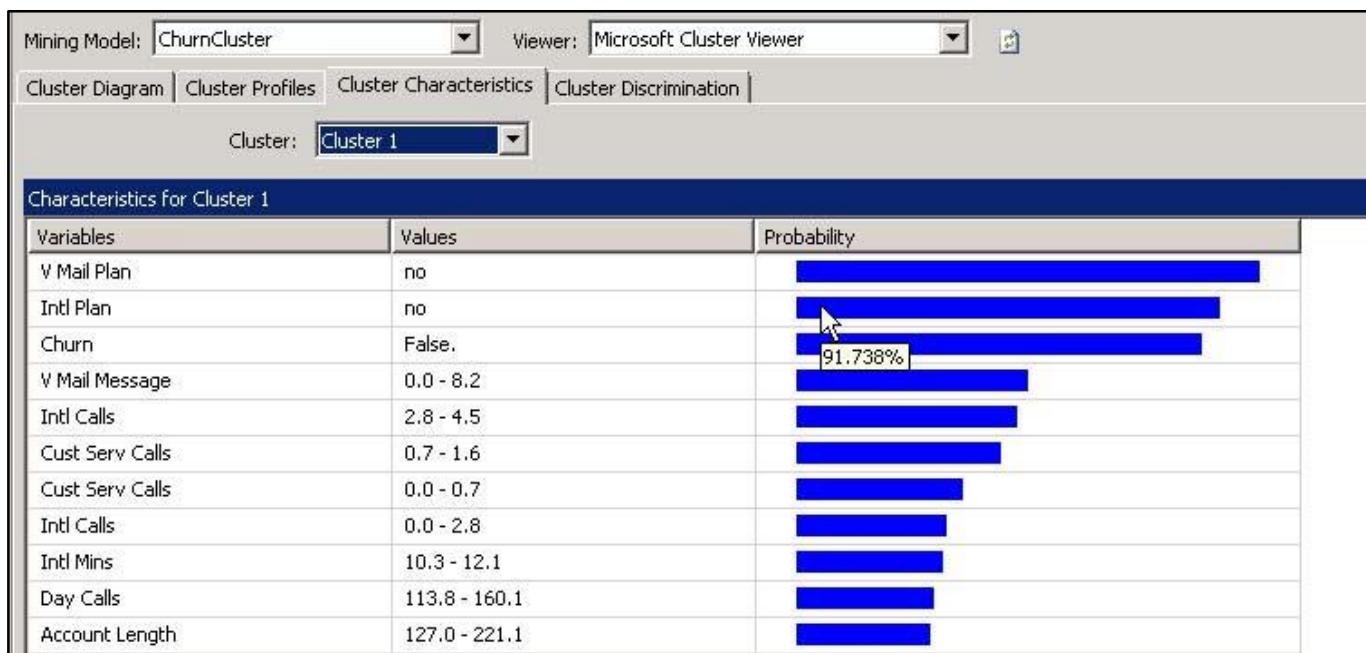
Using the slider bar, one can determine the strongest links are between the different clusters. Click the Cluster Profile sub tab. This visual is very helpful in determining differences in the clusters. Also, for reference, the entire population is presented before the first cluster.



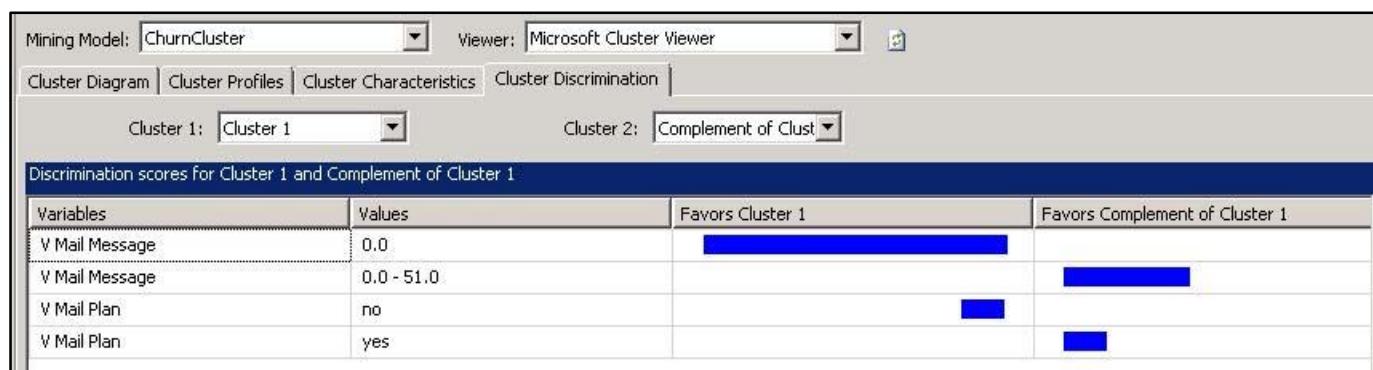
It is not possible to view all the clusters and corresponding attributes in one screen shot—seven attributes are shown in this portion of the cluster profiler. Notice the different visuals for numeric values versus categorical values. Explore the clusters.

Click the Cluster Characteristics sub tab. The Cluster: drop down list box allows the user to select a cluster (All records can also be selected and is the default) for viewing the variables that occur most often in the cluster. Moving the mouse cursor over a bar provides the probability value.

In this cluster, note that almost all of the customers have remained with the telecommunications company—that is they did not churn. Also, note that almost all of these customers also did not have either a voice mail plan or an international plan.



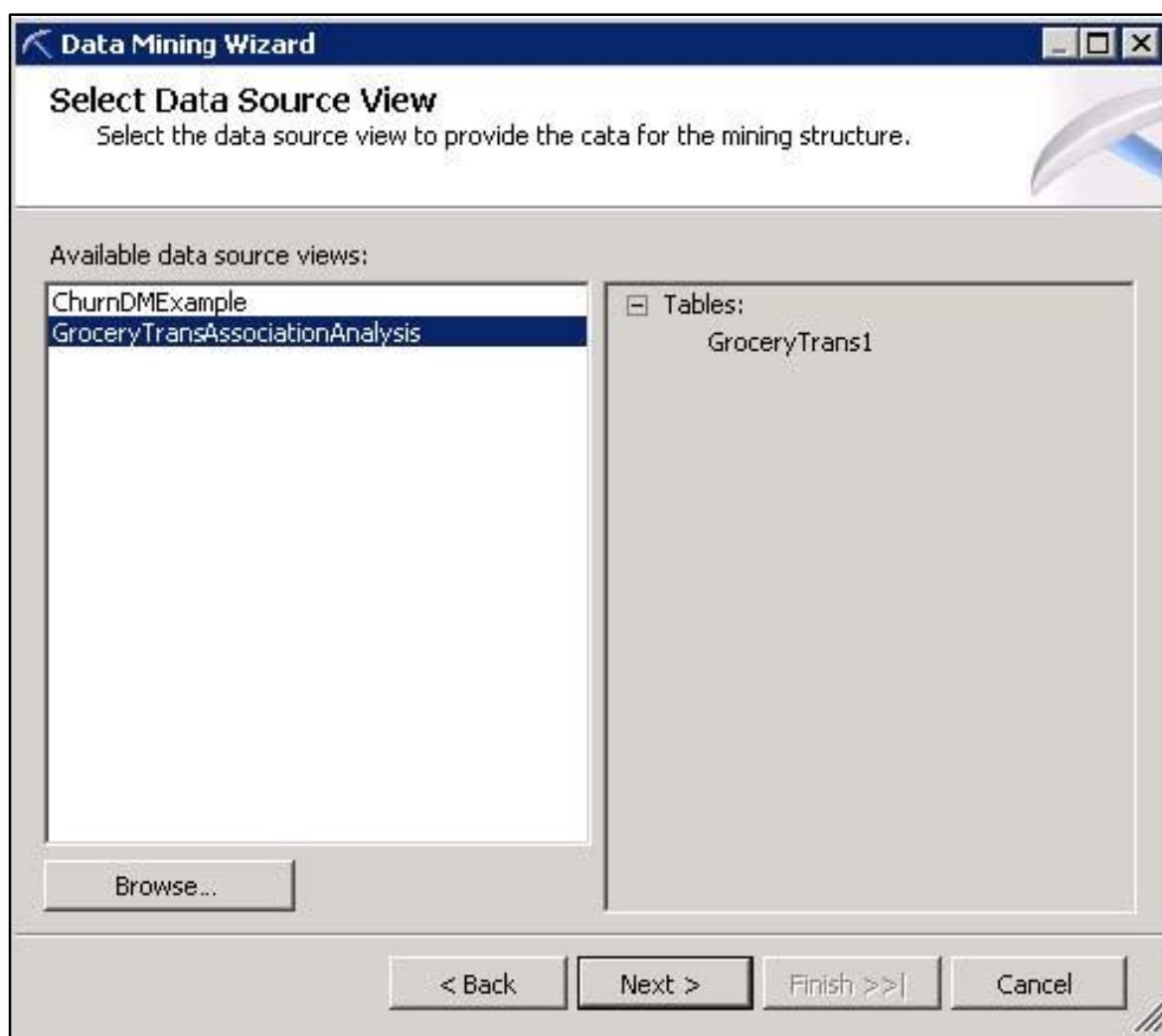
Click the Cluster Discrimination sub tab. This window displays the variables that favor cluster 1 and those that do not favor cluster 1. The user can select clusters from the Cluster 1: drop down list box and compare to the default of Complement of Cluster to specified clusters via selection in the Cluster 2: drop down list box.



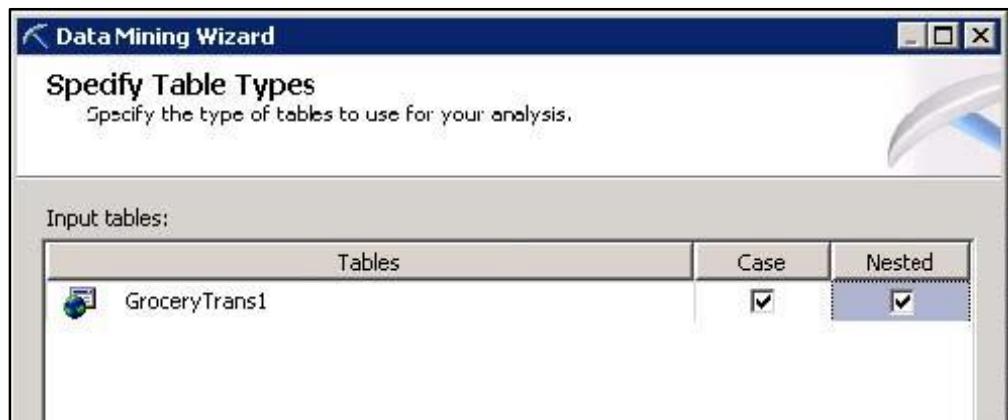
Recommendations: Clustering is very exploratory so you may try different values for the number of customers and also remove the Churn variable and rerun the clustering.

As previously indicated, the Association Analysis will use a different dataset to be more in line with the predominant use of Association Analysis which is for Market Basket Analysis. Thus a new data mining project may be built using the GroceryTrans1 table in the Public\_Datasets\_DM database. Because these steps have already been illustrated, the steps for creating and **Data Source** and a **Data Source View** are not repeated here.

Right click **Mining Structures** in Solution Explorer and create a new mining structure. In this example, **GroceryTrans Association Analysis** is the Data Source View to be used for creating the mining structure.



Select the GroceryTrans1 table and note that in this case, you need to check both the Case and Nested check boxes. Click the Next button.

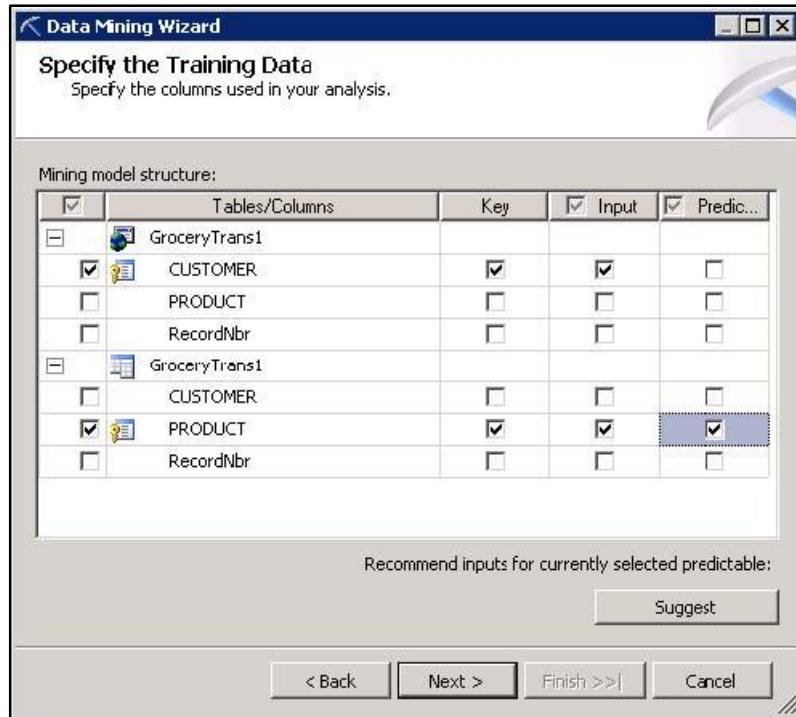


The format of the data is in transactional format and appears as below.

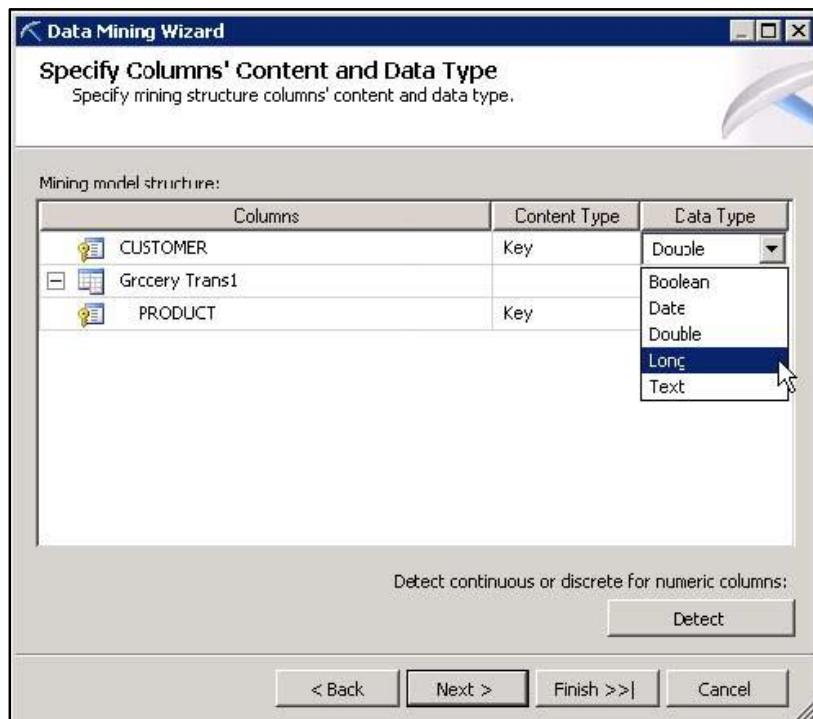
RecordN	CUSTOMER	PRODUCT
br		
1	0	hering
2	0	corned_b
3	0	olives
4	0	ham
5	0	turkey
6	0	bourbon
7	0	ice_crea
8	1	baguette
9	1	soda
10	1	hering
11	1	cracker
12	1	heineken
13	1	olives
14	1	corned_b
15	2	avocado
16	2	cracker

Note that customer is repeated for each item purchased for a single visit – this is referred to as a market basket. Because the customer is repeated, the way this is handled is to have the single table work both as a case and nested table. The Customer will be the key value for the Case portion and the Product will be the key for the nested portion.

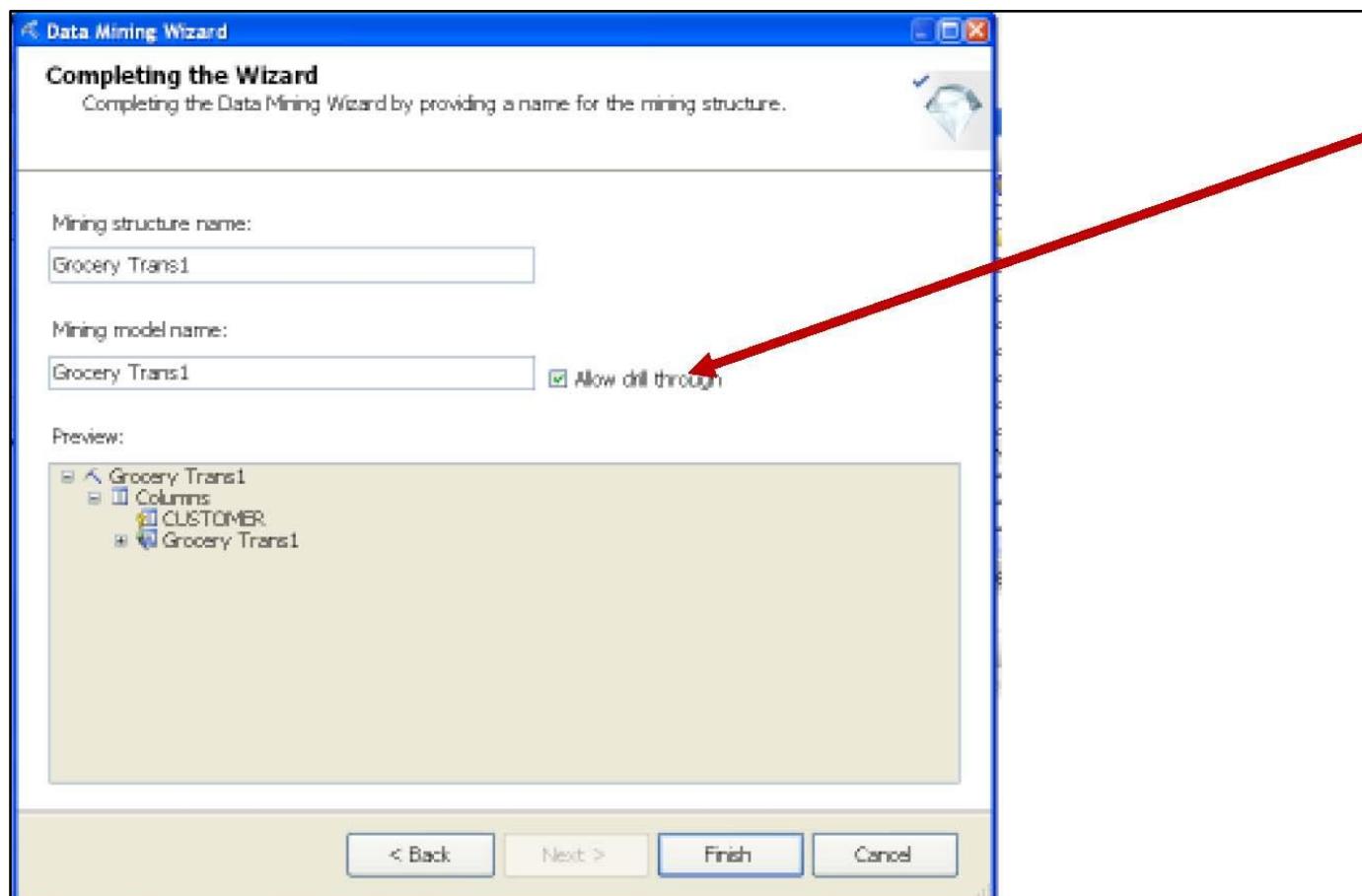
(See exact settings below)



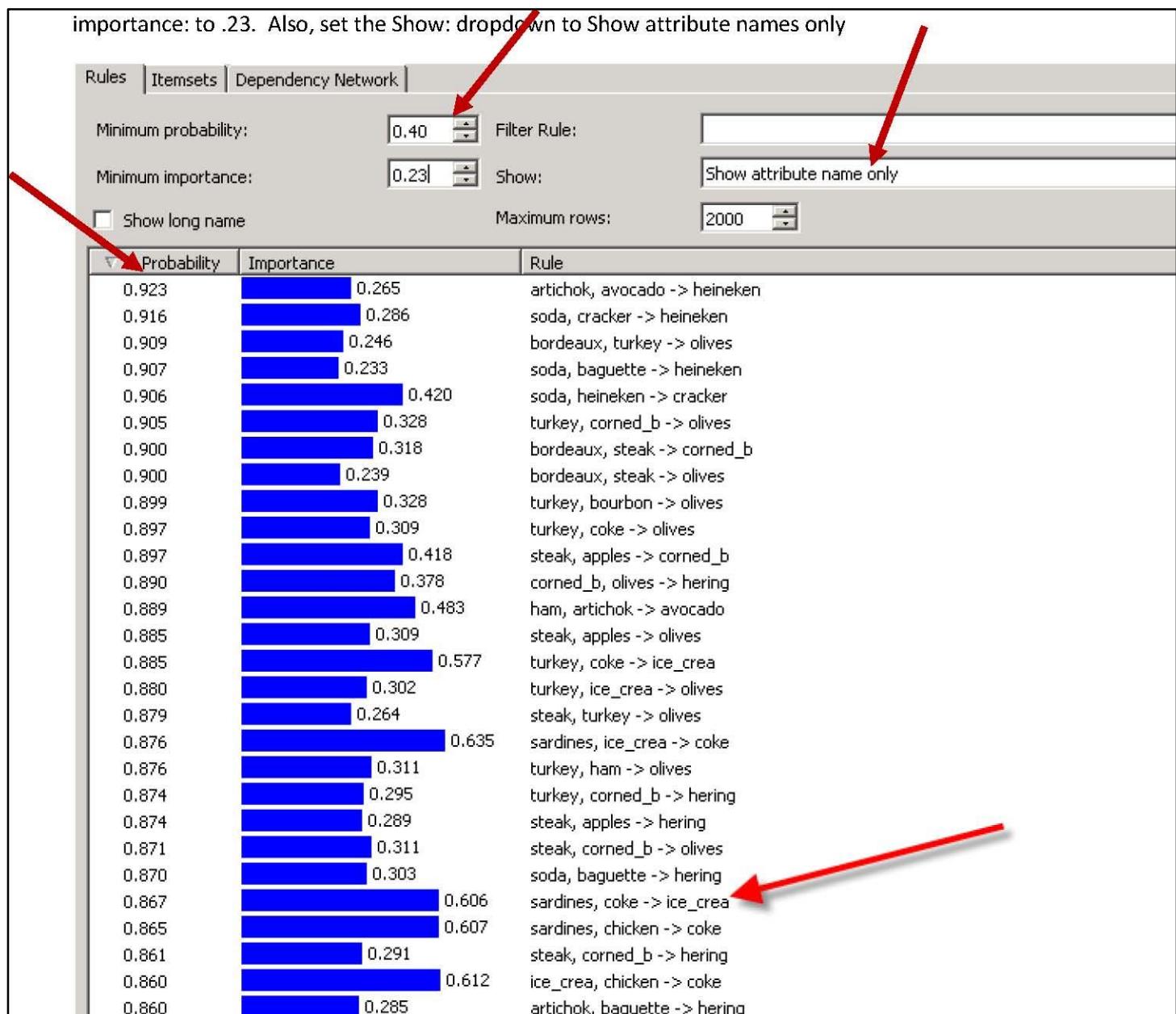
Click the Next button and set the Customer Key to Long (this is not required but the customer number is not decimal so no need for it to be double).



Click the Next button; accept the default of a random test set of 30% and click the Next button. Again, Complete the Data Mining Wizard by providing a Mining Structure name, a Mining Model name and ensuring that the **Allow drill through** check box is checked. See below. Click the Finish button.



Run the model and review the results. The sub tabs for the **Mining Model Viewer** are **Rules**, **Itemsets** and **Dependency Network**. The Rules tab is the default tab with a default minimum probability of .40 displayed and initially sorted by probability—right click the columns and select the desired order. Also, change the Minimum



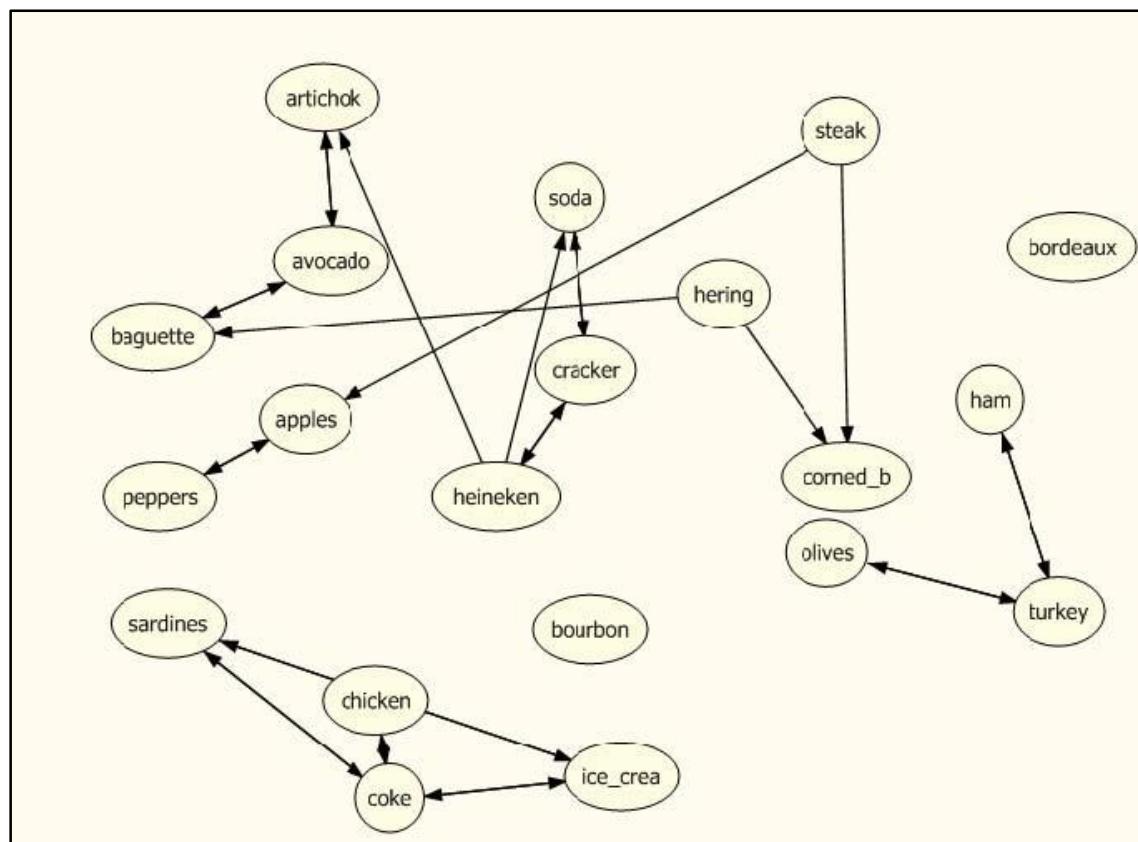
Consider the above rule: sardines, coke-> ice\_crea. It has a fairly high probability and also a fairly high level of Importance. If you have checked to allow drill downs when building the model, Drill down is possible to view customer baskets for this rule. Right click and select **Drill Through**.

Click the Itemsets tab -note that you will probably want to set the Show: drop down box to **Show attribute name only**. Change this to Show attribute name only. Also, note that the number of rows has been set to 2000.

The Itemset shows the Support for each of the products—just a count of how many times the product occurred in the baskets.

Rules	Itemsets	Dependency Network
Minimum support:	1	Filter Itemset:
Minimum itemset size:	0	Show:
Maximum rows:	2000	<input type="checkbox"/> Show long name
Support	Size	Itemset
410	1	heineken
349	1	hering
342	1	cracker
341	1	olives
286	1	corned_b
280	1	bourbon
273	1	baguette

Click the **Dependency Network** sub tab. As with clustering, the slider on the left allows one to investigate the strength of the links between the products—try moving the slider to the top and then to the bottom.



Click on a node and the strength of the links are highlighted. Below, steak has a strong link to apples and corned\_b. Note the legend at the bottom of the screen to indicate the selected product, the products it predicts and the products that predict it.

