# Lesson 1: Preparing the Analysis Services Database (Basic Data Mining Tutorial):

## Creating an Analysis Services Project (Basic Data Mining Tutorial)

### To create an Analysis Services project:

1. Open SQL Server Data Tools (SSDT).
2. On the File menu, point to New, and then select Project.
3. Verify that Business Intelligence Projects is selected in the Project types pane.
4. In the Templates pane, select Analysis Services Multidimensional and Data Mining Project.
5. In the Name box, name the new project BasicDataMining
6. Change File Location
7. Click.

### To change the instance where data mining objects are stored

1. In SQL Server Data Tools (SSDT), on the Project menu, select Properties.
2. On the left side of the Property Pages pane, under Configuration Properties, click Deployment.
3. On the right side of the Property Pages pane, under Target, verify that the Server name is **cssql**. If you are using a different instance, type the name of the instance.
4. Click.

## Creating a Data Source (Basic Data Mining Tutorial)

### To create a data source:

1. In Solution Explorer (on the right), right-click the Data Sources folder and select New Data Source.
2. On the Welcome to the Data Source Wizard page, click Next.
3. On the Select how to define the connection page, click New to add a connection to the database.
4. In the Provider list in Connection Manager, select Native OLE DB\SQL Server Native Client 11.0.
5. In the Server name box, type or select the name of the server on which you installed “**cssql**”.
6. In the Log onto the server group, select Use Windows Authentication.
7. In the Select or enter a database name list, enter “**AdventureWorksDW**” and then click OK.
8. Click Next.
9. On the Impersonation Information page, click Use the username & password, and
   1. then click Next.
   2. On the Completing the Wizard page, notice that, by default, the data source is named Adventure Works DW 2012.
10. Click Finish.
    1. The new data source, Adventure Works DW 2012, appears in the Data Sources folder in Solution Explorer.

## Creating a Data Source View (Basic Data Mining Tutorial)

### To create a data source view:

1. In Solution Explorer (on the right), right-click Data Source Views, and select New Data Source View.
2. On the Welcome to the Data Source View Wizard page, click Next.
3. On the Select a Data Source page, under Relational data sources, select the Adventure Works DW 2012 data source that you created in the last task.
4. Click Next.
   * 1. If you want to create a data source, right-click Data Sources and then click New Data Source to start the Data Source Wizard.
5. On the Select Tables and Views page, select the following objects, and then click the right arrow to include them in the new data source view:
   * ProspectiveBuyer (dbo) - table of prospective bike buyers
   * vTargetMail (dbo) - view of historical data about past bike buyers
6. Click Next.
7. On the Completing the Wizard page, by default the data source view is named Adventure Works DW 2012. Change the name to Targeted Mailing
8. click Finish.
9. The new data source view opens in the Targeted Mailing.dsv [Design] tab

# Lesson 2: Building a Targeted Mailing Structure (Basic Data Mining Tutorial):

## Creating a Targeted Mailing Mining Model Structure (Basic Data Mining Tutorial):

### To create a mining structure for the targeted mailing scenario:

1. In Solution Explorer, right-click Mining Structures and select New Mining Structure to start the Data Mining Wizard.
2. On the Welcome to the Data Mining Wizard page, click Next.
3. On the Select the Definition Method page, verify that From existing relational database or data warehouse is selected, and then click Next.
4. On the Create the Data Mining Structure page, under Which data mining technique do you want to use?, select Microsoft Decision Trees.

**Note**

If you get a warning that no data mining algorithms can be found, the project properties might not be configured correctly. This warning occurs when the project attempts to retrieve a list of data mining algorithms from the Analysis Services server and cannot find the server. By default, SQL Server Data Tools will use **localhost** as the server. If you are using a different instance, or a named instance, you must change the project properties.

1. Click **Next**.
2. On the **Select Data Source View** page, in the **Available data source views** pane, select **Targeted Mailing**. You can click **Browse** to view the tables in the data source view and then click **Close** to return to the wizard.
3. Click **Next**.
4. On the **Specify Table Types** page, select the check box in the **Case** column for vTargetMail to use it as the case table, and then click **Next**. You will use the ProspectiveBuyer table later for testing; ignore it for now.
5. On the **Specify the Training Data** page, you will identify at least one predictable column, one key column, and one input column for your model. Select the check box in the **Predictable** column in the **BikeBuyer** row.

**Note**

Notice the warning at the bottom of the window. You will not be able to navigate to the next page until you select at least one **Input** and one **Predictable** column.

1. Click **Suggest** to open the **Suggest Related Columns** dialog box.

The **Suggest** button is enabled whenever at least one predictable attribute has been selected. The **Suggest Related Columns** dialog box lists the columns that are most closely related to the predictable column, and orders the attributes by their correlation with the predictable attribute. Columns with a significant correlation (confidence greater than 95%) are automatically selected to be included in the model.

Review the suggestions, and then click **Cancel** to ignore the suggestions.

**Note**

If you click **OK**, all listed suggestions will be marked as input columns in the wizard. If you agree with only some of the suggestions, you must change the values manually.

1. Verify that the check box in the **Key** column is selected in the **CustomerKey** row.

**Note**

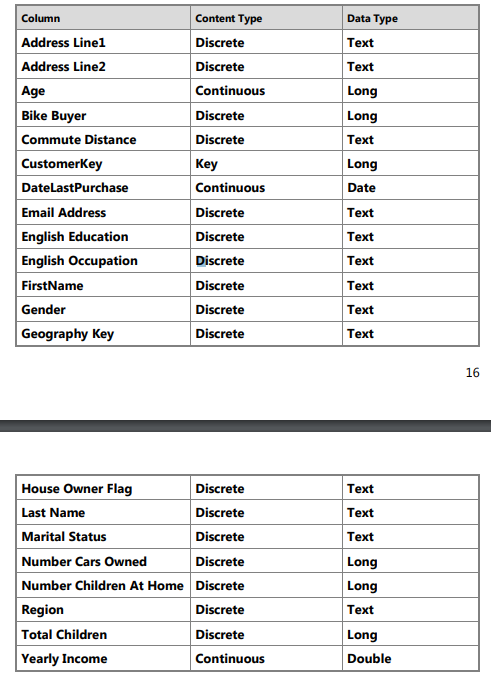
If the source table from the data source view indicates a key, the Data Mining Wizard automatically chooses that column as a key for the model.

1. Select the check boxes in the **Input** column in the following rows. You can check multiple columns by highlighting a range of cells and pressing CTRL while selecting a check box.
2. On the far-left column of the page, select the check boxes in the following rows.
3. Click next.

## Specifying the Data Type and Content Type (Basic Data Mining Tutorial):

### Review and modify content type and data type for each column:

1. On the Specify Columns' Content and Data Type page, click Detect to run an algorithm that determines the default data and content types for each column.
2. Review the entries in the Content Type and Data Type columns and change them if necessary, to make sure that the settings are the same as those listed in the following table.
   * Typically, the wizard will detect numbers and assign an appropriate numeric data type, but there are many scenarios where you might want to handle a number as text instead. For example, the GeographyKey should be handled as text, because it would be inappropriate to perform mathematical operations on this identifier.



Click next

## Specifying a Testing Data Set for the Structure (Basic Data Mining Tutorial):

### Specifying a Testing Set:

1. On the Create Testing Set page, for Percentage of data for testing, leave the default value of 30.
2. For Maximum number of cases in testing data set, type 1000.
3. Click Next.

## Specifying Drill through:

### To name the model and structure and specify drillthrough:

1. On the Completing the Wizard page, in Mining structure name, type Targeted Mailing.
2. In Mining model name, type TM\_Decision\_Tree.
3. Select the Allow drill through check box.
4. Review the Preview pane. Notice that only those columns selected as Key, Input or Predictable are shown. The other columns you selected (e.g., AddressLine1)are not used for building the model but will be available in the underlying structure, and can be queried after the model is processed and deployed.
5. Click Finish.

# Lesson 3: Adding and Processing Models:

## Adding New Models to the Targeted Mailing Structure (Basic Data Mining Tutorial)

### To create a clustering mining mode:

1. Switch to the Mining Models tab in Data Mining Designer in SQL Server Data Tools (SSDT).

Notice that the designer displays two columns, one for the mining structure and one for the TM\_Decision\_Tree mining model, which you created in the previouslesson.

1. Right-click the Structure column and select New Mining Model.
2. In the New Mining Model dialog box, in Model name, type TM\_Clustering.
3. In Algorithm name, select Microsoft Clustering.
4. Click .

### To create a Naive Bayes mining model:

1. In the Mining Models tab of Data Mining Designer, right-click the Structure

column, and select New Mining Model.

2. In the New Mining Model dialog box, under Model name, type

TM\_NaiveBayes.

3. In Algorithm name, select Microsoft Naive Bayes, then click OK.

A message appears stating that the Microsoft Naive Bayes algorithm does not

support the Age and Yearly Income columns, which are continuous.

1. Click Yes to acknowledge the message and continue.

## Processing Models in the Targeted Mailing Structure (Basic Data Mining Tutorial) Before you can browse

### To set the Holdout Seed;

1. Click on the Mining Structure tab or the Mining Models tab in Data Mining Designer in SQL Server Data Tools (SSDT). Targeted Mailing MiningStructure displays in the Properties pane.
2. Ensure that the Properties pane is open by pressing F4.
3. Ensure that CacheMode is set to KeepTrainingCases.
4. Enter 12 for HoldoutSeed.

## Deploying and Processing the Models:

### To deploy the project and process all the mining models

1. In the Mining Model menu, select Process Mining Structure and All Models.

If you made changes to the structure, you will be prompted to build and deploy the project before processing the models. Click Yes.

1. Click Run in the Processing Mining Structure - Targeted Mailing dialog box.

The Process Progress dialog box opens to display the details of model processing. Model processing might take some time, depending on your computer.

1. Click Close in the Process Progress dialog box after the models have completed processing.
2. Click Close in the Processing Mining Structure - <structure> dialog box.

# Lesson 4: Exploring the Targeted Mailing Models (Basic Data Mining Tutorial):

## To explore the model in the Decision Tree tab:

1. Select the **Mining Model Viewer** tab in **Data Mining Designer**.
   1. By default, the designer opens to the first model that was added to the structure -- in this case, **TM\_Decision\_Tree**.
2. Use the magnifying glass buttons to adjust the size of the tree display.
   1. By default, the Microsoft Tree Viewer shows only the first three levels of the tree. If the tree contains fewer than three levels, the viewer shows only the existing levels. You can view more levels by using the **Show Level** slider or the **Default Expansion** list.
3. Slide **Show Level** to the fourth bar.
4. Change the **Background** value to **1**.
   1. By changing the **Background** setting, you can quickly see the number of cases in each node that have the target value of **1** for [Bike Buyer]. Remember that in this particular scenario, each case represents a customer. The value **1** indicates that the customer previously purchased a bike; the value **0** indicates that the customer has not purchased a bike. The darker the shading of the node, the higher the percentage of cases in the node that have the target value.
5. Place your cursor over the node labeled **All**. An tooltip will display the following information:

* Total number of cases
* Number of non bike buyer cases
* Number of bike buyer cases
* Number of cases with missing values for [Bike Buyer]

Alternately, place your cursor over any node in the tree to see the condition that is required to reach that node from the node that comes before it. You can also view this same information in the **Mining Legend**.

1. Click on the node for **Age >=34 and < 41**. The histogram is displayed as a thin horizontal bar across the node and represents the distribution of customers in this age range who previously did (pink) and did not (blue) purchase a bike. The Viewer shows us that customers between the ages of 34 and 40 with one or no cars are likely to purchase a bike. Taking it one step further, we find that the likelihood to purchase a bike increases if the customer is actually age 38 to 40.

### To drill through to case data

1. Right-click a node, and select Drill Through then Model Columns Only.

The details for each training case are displayed in spreadsheet format. These details come from the vTargetMail view that you selected as the case table when building the mining structure.

1. Right-click a node, and select Drill Through then Model and Structure

### To explore the model in the Dependency Network tab

1. Click the Bike Buyer node to identify its dependencies.

The center node for the dependency network, Bike Buyer, represents the predictable attribute in the mining model. The pink shading indicates that all of the attributes have an effect on bike buying.

1. Adjust the All Links slider to identify the most influential attribute.

As you lower the slider, only the attributes that have the greatest effect on the [Bike Buyer] column remain. By adjusting the slider, you can discover that age and region are the greatest factors in predicting whether someone is a bike buyer

## Exploring the Clustering Model (Basic Data Mining Tutorial)

### To explore the model in the Cluster Diagram tab:

1. Use the Mining Model list at the top of the Mining Model Viewer tab to switch to the TM\_Clustering model.
2. In the Viewer list, select Microsoft Cluster Viewer.
3. In the Shading Variable box, select Bike Buyer.

The default variable is Population, but you can change this to any attribute in the model, to discover which clusters contain members that have the attributes you want.

1. Select 1 in the State box to explore those cases where a bike was purchased.

The Density legend describes the density of the attribute state pair selected in the Shading Variable and the State. In this example it tells us that the cluster with the darkest shading has the highest percentage of bike buyers.

1. Pause your mouse over the cluster with the darkest shading.

A tooltip displays the percentage of cases that have the attribute, Bike Buyer = 1.

1. Select the cluster that has the highest density, right-click the cluster, select Rename Cluster and type Bike Buyers High for later identification. Click .
2. Find the cluster that has the lightest shading (and the lowest density). Right-click the cluster, select Rename Cluster and type Bike Buyers Low. Click .
3. Click the Bike Buyers High cluster and drag it to an area of the pane that will give you a clear view of its connections to the other clusters.

When you select a cluster, the lines that connect this cluster to other clusters are highlighted, so that you can easily see all the relationships for this cluster. When the cluster is not selected, you can tell by the darkness of the lines how strong the relationships are amongst all the clusters in the diagram. If the shading is light or non-existent, the clusters are not very similar.

1. Use the slider to the left of the network, to filter out the weaker links and find the clusters with the closest relationships. The Adventure Works Cycles marketing department might want to combine similar clusters together when determining the best method for delivering the targeted mailing.

### To explore the model in the Cluster Profiles tab:

1. Set Histogram bars to 5. In our model, 5 is the maximum number of states for any one variable.
2. If the Mining Legend blocks the display of the Attribute profiles, move it out of the way.
3. Select the Bike Buyers High column and drag it to the right of the Population column.
4. Select the Bike Buyers Low column and drag it to the right of the Bike Buyers High column.
5. Click the Bike Buyers High column. The Variables column is sorted in order of importance for that cluster. Scroll through the column and review characteristics of the Bike Buyer High cluster. For example, they are more likely to have a short commute.
6. Double-click the Age cell in the Bike Buyers High column. The Mining Legend displays a more detailed view and you can see the age range of these customers as well as the mean age
7. Right-click the Bike Buyers Low column and select Hide Column.

### To explore the model in the Cluster Discrimination tab:

1. In the Cluster 1 box, select Bike Buyers High.
2. In the Cluster 2 box, select Bike Buyers Low.
3. Click Variables to sort alphabetically.

Some of the more substantial differences among the customers in the Bike Buyers Low and Bike Buyers High clusters include age, car ownership, number of children, and region.

## Exploring the Naive Bayes Model (Basic Data Mining Tutorial)

### To explore the model in the Dependency Network tab

1. Use the Mining Model list at the top of the Mining Model Viewer tab to switch to the TM\_NaiveBayes model.
2. Use the Viewer list to switch to Microsoft Naive Bayes Viewer.
3. Click the Bike Buyer node to identify its dependencies. The pink shading indicates that all of the attributes have an effect on bike buying.
4. Adjust the slider to identify the most influential attribute.

As you lower the slider, only the attributes that have the greatest effect on the [Bike Buyer] column remain. By adjusting the slider, you can discover that a few of the most influential attributes are: number of cars owned, commute distance, and total number of children.

### To explore the model in the Attribute Profiles tab

1. In the Predictable box, verify that Bike Buyer is selected.
2. If the Mining Legend is blocking display of the Attribute profiles, move it out of the way.
3. In the Histogram bars box, select 5.

In our model, 5 is the maximum number of states for any one variable. The attributes that affect the state of this predictable attribute are listed together with the values of each state of the input attributes and their distributions in each state of the predictable attribute.

1. In the Attributes column, find Number Cars Owned. Notice the differences in the histograms for bike buyers (column labeled 1) and non-buyers (column labeled 0). A person with zero or one car is much more likely to buy a bike.
2. Double-click the Number Cars Owned cell in the bike buyer (column labeled 1) Column

# Lesson 5: Testing Models (Basic Data Mining Tutorial):

## Choosing the Input Data:

### To select the data set:

1. Switch to the Mining Accuracy Chart tab in Data Mining Designer in SQL Server Data Tools (SSDT) and select the Input Selection tab.
2. In the Select data set to be used for Accuracy Chart group box (**bottom of screen**), select Use mining structure test cases to test your models by using the testing data that you set aside when you created the mining structure.

### To show the lift of the models:

1. On the Input Selection tab of Data Mining Designer, under Select predictable mining model columns to show in the lift chart, select the checkbox for Synchronize Prediction Columns and Values.
2. In the Predictable Column Name column, verify that Bike Buyer is selected for each model.
3. In the Show column, select each of the models. By default, all the models in the mining structure are selected. You can decide not to include a model, but for this tutorial leave all the models selected.
4. In the Predict Value column, select 1. The same value is automatically filled in for each model that has the same predictable column.
5. Select the Lift Chart tab to display the lift chart. When you click the tab, a prediction query runs against the server and database for the mining structure and the input table or test data. The results are plotted on the graph.

When you enter a **Predict Value**, the lift chart plots a Random Guess Model as well as an Ideal Model. The mining models you created will fall between these two extremes; between a random guess and a perfect prediction. Any improvement from the random guess is considered to be *lift*.

1. Use the legend to locate the colored lines representing the Ideal Model and the Random Guess Model. You'll notice that the TM\_Decision\_Tree model provides the greatest lift, outperforming both the Clustering and Naive Bayes models.

## Testing a Filtered Model (Basic Data Mining Tutorial):

### To copy the Decision Tree Model:

1. Click the **Mining Models** tab.
2. Right click the v Targeted Mail model, and select New Mining Model.
3. In the Model name field, type TM\_Decision\_Tree\_Male.
4. Click OK.

### To create a case filter on a mining model:

1. Right-click the **TM\_Decision\_Tree\_Male.**
2. Set Model filter
3. In the **Model Filter** dialog box, click the top row in the grid, in the **Mining Structure Column** text box. The drop-down list displays only the names of the columns in that table.
4. In the Mining Structure Column text box, select **Gender**. The icon at the left side of the text box changes to indicate that the selected item is a table or a column.
5. Click the **Operator** text box and select the equal (=) operator from the list.
6. Click the **Value** text box, and type **M**.
7. Click the next row in the grid.
8. Click **OK** to close the **Model Filter** dialog box. The filter displays in the **Properties** window. Alternately, you can launch the **Model Filter** dialog from the **Properties** window. 8. Repeat the above steps, but this time name the model **TM\_Decision\_Tree\_Female** and type **F** in the **Value** text box.

You now have two new models displayed in the **Mining Models** tab.

**Process the Filtered Models:**

Models cannot be used until they have been deployed and processed. For more information on processing models

### To process the filtered model:

1. Right-click the **TM\_Decision\_Tree\_Male** model and select **Process Mining Structure and all Model**s
2. Click **Run** to process the new models.
3. After processing is complete, click **Close** on both processing windows.

**Evaluate the Results**

### To explore the filtered models:

1. Select the Mining Model Viewer tab in Data Mining Designer.
2. In the Mining Model box, select TM\_Decision\_Tree\_Male.
3. Slide Show Level to 3.
4. Change the Background value to 1.
5. Place your cursor over the node labeled All to see the number of bike buyers versus non-bike buyers.
6. Repeat steps 1 - 5 for TM\_Decision\_Tree\_Female.
7. Explore the results for the TM\_Decision\_Tree and the models filtered for gender. Compared to all bike buyers, male and female bike buyers share some of the same characteristics as the unfiltered bike buyers but all three have interesting differences as well. This is useful information that Adventure Works Cycles can use to develop their marketing campaign.

### To test the lift of the filtered models:

1. Switch to the **Mining Accuracy Chart** tab in Data Mining Designer in SQL Server Data Tools (SSDT) and select the **Input Selection** tab.
2. In the **Select data set to be used for Accuracy Chart** group box, select **Use mining structure test cases**.
3. On the **Input Selection** tab of Data Mining Designer, under **Select predictable mining model columns to show in the lift chart**, select the checkbox for **Synchronize Prediction Columns and Values**.
4. In the **Predictable Column Name** column, verify that **Bike Buyer** is selected for each model.
5. In the **Show** column, select each of the models.
6. In the **Predict Value** column, select **1**.
7. Select the **Lift Chart** tab to display the lift chart.

You will now notice that all three Decision Tree models provide significant lift compared to the random guess model, and also outperform the Clustering and

Naive-Bayes models.

**Intermediate Data Mining Tutorial (Analysis Services - Data Mining)**

**Lesson 1: Creating the Intermediate Data Mining Solution (Intermediate Data Mining Tutorial):**

**To create a new Analysis Services project for this tutorial:**

1. Open SQL Server Data Tools (SSDT).

2. On the **File** menu, point to **New**, and then click **Project**.

3. Select **Analysis Services Multidimensional and Data Mining Project** from the **Installed Templates** pane.

4. In the **Name** box, name the new project **DM Intermediate**.

5. Click .

**Lesson 2: Building a Forecasting Scenario (Intermediate Data Mining Tutorial)**

**To add a data source view**

1. In Solution Explorer, right-click **Data Source Views**, and then select **New Data Source View**.

2. On the **Welcome to the Data Source View Wizard** page, click **Next**.

3. On the **Select a Data Source** page, under **Relational data sources**, select the data source. Click **Next**. **Note**

4. On the **Select Tables and Views** page, select the table, vTimeSeries (dbo), and then click the right arrow to add it to the data source view.

5. Click **Next**.

6. On the **Completing the Wizard** page, by default the data source view is named Adventure Works DW Multidimensional 2012 . Change the name to **SalesByRegion**, and then click **Finish**.

Data Source View Designer opens and the **SalesByRegion** data source view appears.

**Working with the Data Source View**

After you have created the data source view, you can explore the data in the following ways:

• Right-click the table vTimeSeries in the designer, and select **Explore Data** to open the selected table in a grid.

• Click **Sampling options** and then use the **Data Exploration Options** dialog box to change the sampling method. Click **Refresh** to load data in the table using the new option settings. For example, you could specify the number of rows to output in the sample, or choose the top rows.

• Right-click the table vTimeSeries and select **Properties** to assign a new name to the table. You can also select individual columns from the data source view, and the modify the column properties.

• Click anywhere in the data source view design area to create a new query and assign a name to it, to create relationships between tables, or to change the layout of the design area.

• Right-click a table and select **New Named Calculation** to create derived columns, including aggregations. You can also add new tables and views from the data source in this view. In the next task, you will explore the time series data and determine the best column to use as the time series identifier. You will also learn how to handle gaps in time series data

**To identify the time key for the forecasting model:**

1. In the pane, **SalesByRegion.dsv [Design]**, (right click open first) right-click the table vTimeSeries, and then select **Explore Data**. A new tab opens, titled **Explore vTimeSeries Table**.
2. On the **Table** tab, review the data that is used in the TimeIndex and Reporting Date columns. Both are sequences with unique values and can both be used as the time series key; however, the data types of the columns are different. The Microsoft Time Series algorithm does not require a **datetime** data type, only that the values be distinct and ordered. Therefore, either column can be used as the time key for the forecasting model.
3. In the data source view design surface, select the column, Reporting Date and select **Properties**. Next, click the column TimeIndex and select **Properties**.

**To set the key in the data source view:**

1. In the pane **SalesByRegion.dsv**, select the vTimeSeries table.

2. Right-click the column, Reporting Date, and select **Set Logical Primary Key**.

**To create a forecasting mining structure:**

1. In Solution Explorer in SQL Server Data Tools (SSDT), right-click **Mining Structures** and select **New Mining Structure**.
2. On the **Welcome to the Data Mining Wizard** page, click **Next**.
3. On the **Select the Definition Method** page, verify that **From existing relational database or data warehouse** is selected, and then click **Next**.
4. On the **Create the Data Mining Structure** page, under **Which data mining technique do you want to use?**, select **Microsoft Time Series**, and then click **Next**.
5. On the **Select Data Source View** page, under **Available data source views**, select **SalesByRegion**.
6. Click **Next**.
7. On the **Specify Table Types** page, ensure that the check box in the **Case** column for the vTimeSeries table is selected, and then click **Next**.
8. On the **Specify the Training Data** page, select the check boxes in the **Key** column for the ModelRegion and ReportingDate columns. ReportingDate should be selected by default, because you specified this column as the logical primary key when you created the data source view. By adding ModelRegion as a second key, you are telling the algorithm to create a separate time series for each combination of model and region listed in this field.
9. Select the check boxes in the **Input** and **Predictable** columns for the Quantity, column, and then click **Next**. By selecting **Predictable**, you indicate that you want to create forecasts on the data in this column. However, because you want to base the forecasts on past data, you must also add the column as an input.
10. On the page **Specify Columns' Content and Data Type**, review the selections. The ModelRegion column is designated as a **Key** column and the ReportingDate column is automatically designated as a **Key Time** column. You can have only one of each type of key.
11. Click **Next**.
12. On the **Completing the Wizard** page, for **Mining structure name**, type

**Modifying the Forecasting Structure (Intermediate Data Mining Tutorial):**

**To add the Amount column to the Forecasting mining structure**

1. On the **Mining Structure** tab of Data Mining Designer, in the **Data Source View** pane, select the Amount column in the vTimeSeries table.
2. Drag the Amount column from the **Data Source View** pane into the list of columns for the **Forecasting** structure. The Amount column is now included in the **Forecasting** mining structure.