Weekly Exercise 4

NOTE: In a SSDT Analysis Services project, you will need to analyze the data from a data source (usually a relational database), then store the data mining objects, such as algorithms used and models created, in an Analysis Services database, which is a special database for OLAP cubes and data mining objects. In this tutorial, the data source is the relational database: AdventureWorksDW2012 in the relational database server Microsoft SQL Server. The Analysis Services database is created in the Microsoft Analysis Services Management System during the tutorial. We use the same name for the Analysis Services project and database in this format DM-YourWindowsLoginName, such as DM-chenx. The names of the project name and database are not required to be the same. This tutorial consists of five lessons.

1. **Lesson 1-createing an Analysis Services project and preparing the Analysis Services database with the name DM-YourWindowsLoginName.**
2. **Lesson 2-building a target mailing structure, which includes data items and one mining model (in this tutorial, a Decision Tree model for classification).**
3. **Lesson 3-adding more mining models and processing the models for prediction (in this tutorial, a clustering model for segmentation and a Naïve Bayes model for classification).**
4. **Lesson 4- Exploring the Targeted Mailing Models. Read the explanations in this lesson carefully to understand what each model can provide for prediction.**

Lesson 1: Creating an Analysis Services project and preparing an Analysis Services Database

1. To create an Analysis Services project in SSDT

1. Open Visual Studio (or SQL Server Data Tools (SSDT) if you have it installed as a standalone application).
2. On the **File** menu, point to **New**, and then select **Project**.
3. Select Multidimensional under Analysis Services folder, then in the **Templates** pane, select **Analysis Services Multidimensional and Data Mining Project**.
4. In the **Name** box, name the new project DM-YourWindowsLoginName (by default, the Analysis Services database name on the server will be named after the project name).
5. Click Browse… button, then choose where you want to store the project. Click **OK**.

2. To change the database where data mining objects are stored

1. Right click DM-YourWindowsLoginName in the Solution Explorer window, then 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 **misbi.cbe.wwu.edu\multi** and the database name is DM-YourWindowsLoginName, then click **OK**.

3. To create a data source

1. In **Solution Explorer**, 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, select the radio button of Create a data source based on an existing or new connection, then click **New** to add a connection to the **AdventureWorksDW2012** 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 misbi.cbe.wwu.edu\multi
6. In the **Log onto the server** group, select SQL Server Authentication
7. Enter your user name and password for the authentication. The user name is in the format of QYY_YourWindowsLogin, where Q represents the quarter (S for Spring, W for Winter, F for Fall, and SUM for Summer) and YY the year. For example, S19 is used for the Spring Quarter of 2019. The password is your Western ID starting with W.
8. In the **Select or enter a database name** list, select **AdventureWorksDW2012**and then click **OK**. Click **Next**.
9. On the **Impersonation Information** page, click **Use the service account**, and then click **Next**.
   On the **Completing the Wizard** page, notice that, by default, the data source is named Adventure Works DW2012.
10. Click **Finish**.
    The new data source, Adventure Works DW 2012, appears in the **Data Sources** folder in Solution Explorer.

4. To create a data source view

A data source view is built on a data source and defines a subset of the data, which you can then use in your mining structures. You can also use the data source view to add columns, create calculated columns and aggregates, and add named views. By using data source views, you can select the data that relates to your project, establish relationships between tables, and modify the structure of the data, without modifying the original data source.

1. In **Solution Explorer**, right-click **Data Source Views**, and select **New Data Source View**.
2. On the **Welcome to the Data Source View Wizard** page, click **Next**.

This document is adapted from Microsoft Tutorial. Credits go to the Microsoft.

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. Click **Next**.
   Note If you want to create a data source, right-click **Data Sources** and then click **New Data Source** to start the Data Source Wizard.
4. 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
5. Click **Next**.
6. On the **Completing the Wizard** page, by default the data source view is named Adventure Works DW 2012. Change the name to YourWindowsLoginName-Targeted Mailing, and then click **Finish**.
   Double click the new data source view YourWindowsLoginName-Targeted Mailing.dsv under Data Source Views folder in Solution Explorer window, the view opens in the **YourWindowsLoginName-Targeted Mailing.dsv [Design]** tab.

Lesson 2: Building a Target Mailing Structure

The Marketing department of Adventure Works Cycles wants to increase sales by targeting specific customers for a mailing campaign. The company's database contains a list of past customers (in vTargetMail (dbo)) and a list of potential new customers (in ProspectiveBuyer (dbo)). By investigating the attributes of previous customers, the company hopes to discover patterns that they can then apply to potential customers. For example, they might use past trends to predict which potential customers are most likely to purchase a bike from Adventure Works Cycles, or create customer segments for future marketing campaigns. In this lesson, you will use the **Data Mining Wizard** to create the targeted mailing structure. After you complete the tasks in this lesson, you will have a mining structure with a single model.

1. 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**. Click **Next**.
   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. Right click the DM-YourWindowsLoginName in the Solutin Explorer window, select Properties. Click the Deployment tab on the DM-YourWindowsLoginName, make sure the server is misbi.cbe.wwu.edu\multi.
5. On the **Select Data Source View** page, in the **Available data source views** pane, select **YourWindowsLoginName-Targeted Mailing**. You can click **Browse** to view the tables in the data source view and then click **Close** to return to the wizard. Click **Next**.
6. On the **Specify Table Types** page, select the check box in the **Case** column for vTargetMail to use it as the case table (for training the model), and then click **Next**. You will use the ProspectiveBuyer table later for testing; ignore it for now.
7. 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.
8. 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.
9. 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.
10. 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.
    - **Age**
    - **CommuteDistance**
    - **EnglishEducation**
    - **EnglishOccupation**
    - **Gender**
    - **GeographyKey**
    - **HouseOwnerFlag**
    - **MaritalStatus**
    - **NumberCarsOwned**
    - **NumberChildrenAtHome**

- **Region**
- **TotalChildren**
- **YearlyIncome**

11. On the far left column of the page, select the check boxes in the following rows.
    - **AddressLine1**
    - **AddressLine2**
    - **DateFirstPurchase**
    - **EmailAddress**
    - **FirstName**
    - **LastName**

    Ensure that these rows have checks only in the left column. These columns will be added to your structure but will not be included in the model. However, after the model is built, they will be available for drillthrough and testing. For more information about drillthrough, see Drillthrough Queries (Data Mining). Click **Next**.

2. Review and modify content type and data type for each column

Now that you have selected which columns to use for building your structure and training your models, make any necessary changes to the default data and content types that are set by the wizard.

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.

Column	Content Type	Data Type
Address Line1	**Discrete**	**Text**
Address Line2	**Discrete**	**Text**
Age	**Continuous**	**Long**
Bike Buyer	**Discrete**	**Long**
Commute Distance	**Discrete**	**Text**
CustomerKey	**Key**	**Long**
DateLastPurchase	**Continuous**	**Date**
Email Address	**Discrete**	**Text**
English Education	**Discrete**	**Text**
English Occupation	**Discrete**	**Text**
FirstName	**Discrete**	**Text**
Gender	**Discrete**	**Text**
Geography Key	**Discrete**	**Text**
House Owner Flag	**Discrete**	**Text**
Last Name	**Discrete**	**Text**
Marital Status	**Discrete**	**Text**
Number Cars Owned	**Discrete**	**Long**
Number Children At Home	**Discrete**	**Long**
Region	**Discrete**	**Text**
Total Children	**Discrete**	**Long**
Yearly Income	**Continuous**	**Double**

3. Click **Next**.

3. Specifying a Testing Set

Separating data into training and testing sets when you create a mining structure makes it possible to easily assess the accuracy of the mining models that you create later.

To specify the 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. Click **Next**.

4. Specifying Drillthrough

Drillthrough can be enabled on models and on structures. The checkbox in this dialog box enables drillthrough on the named model. After the model has been processed, you will be able to retrieve detailed information from the training data that were used to create the model. If the underlying mining structure has also been configured to allow drillthrough, you can retrieve detailed information from both the model cases and

the mining structure, including columns that were not included in the mining model. For more information, see Drillthrough Queries (Data Mining).

To name the model and structure and specify drillthrough

1. On the **Completing the Wizard** page, in **Mining structure name**, type YourWindowsLoginName-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. Click **Finish**.

Lesson 3: Adding and Processing Models

The mining structure that you created in the previous lesson contains a single mining model that is based on the Microsoft Decision Trees algorithm. You can use this model to identify customers for the targeted mailing campaign. However, to ensure that your analysis is thorough, it is a common practice to create related models using different algorithms and compare their results. That way you can get different insights as well. Therefore, you will create two additional models, then process and deploy the models. In this lesson, you will create two new mining models: clustering and Naïve Bayes mining models that will suggest the most likely customers from a list of potential customers.

1. To create a clustering mining model

1. Switch to the **Mining Models** tab in Data Mining Designer in Visual Studio (or 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 previous lesson.
2. Right-click the **Structure** column and select **New Mining Model**.
3. In the **New Mining Model** dialog box, in **Model name**, type TM_Clustering.
4. In **Algorithm name**, select **Microsoft Clustering**. Click **OK**.

The new model now appears in the **Mining Models** tab of Data Mining Designer. This model, built with the Microsoft Clustering algorithm, groups customers with similar characteristics into clusters and predicts bike buying for each cluster. Although you can modify the column usage and properties for the new model, no changes to the TM_Clustering model are necessary for this tutorial.

2. To create a Naive Bayes mining model

1. In the **Mining Models** tab of Data Mining Designer, right-clickthe **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 (Remember: Naïve Bayes classifier can only take categorical variables. we could convert the values of these two variables into categorical values, then they can be used in this model). Click **Yes** to acknowledge the message and continue.

A new model appears in the **Mining Models** tab of Data Mining Designer. Although you can modify the column usage and properties for all the models in this tab, no changes to the TM_NaiveBayes model are necessary for this tutorial.

Before you can browse or work with the mining models that you have created, you must deploy the Analysis Services project and process the mining structure and mining models.

- *Deploying* sends the project to a server and creates any objects in that project on the server.
- *Processing* populates Analysis Services objects with data from relational data sources.

Models cannot be used until they have been deployed and processed. Also, when you make any changes to the model, such as adding new data, you must redeploy and reprocess the models.

3. To ensure Consistency with HoldoutSeed

When you deploy a project and process the structure and models, individual rows in your data structure are assigned to either the training set or testing set based on a numerical seed value. By default, the numerical seed value is computed based on attributes of the data structure. However, if you ever change some aspects of your model, the seed value would change, leading to subtly different results. Therefore, in order to ensure that your results are the same as described here, we will arbitrarily assign a fixed *holdout seed* of 12. The holdout seed is used to initialize the sampling algorithm, and ensures that the data is partitioned in roughly the same way for all mining structures and their models. This value does not affect the number of cases in the training set; it simply ensures that the same partitioning method will be used each time you build the model. For more information on holdout seed, see Training and Testing Data Sets.

This document is adapted from Microsoft Tutorial. Credits go to the Microsoft.

To set the Holdout Seed

1. Click on the **Mining Structure** tab or the **Mining Models** tab in Data Mining Designer in Visual Studio (or SSDT). **YourWindowsLoginName-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**.

4. Deploying and Processing the Models

In Data Mining Designer, you can decide which objects to process, depending on the scope of changes you've made to your model or the underlying data. For this task, because the data and the models are new, you will process the structure and all the models at the same time.

To deploy the project and process all the mining models

1. In the **Mining Model** menu, select **Process Mining Structure and All Models**. You can also click the button  on the Mining Models tab to process the 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**.
2. Click **Run** in the **Processing Mining Structure – YourWindowsLoginName-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.
3. Click **Close** in the **Process Progress** dialog box after the models have completed processing.
4. Click **Close** in the **Processing Mining Structure - <structure>** dialog box.

Lesson 4: Exploring the Targeted Mailing Models

After the models in your project have been processed, you can explore them to look for interesting trends. Because patterns can be complex and difficult simply by looking at numbers, SQL Server Data Mining provides some visual tools that help you investigate the data and understand the rules and relationships that the algorithms have discovered within the data. You can also use a variety of accuracy tests to validate your dataset or discover which model performs best before you deploy it. When you use SQL Server Data Tools (SSDT) to explore your models, each model you created is listed in the **Mining Model Viewer** tab in Data Mining Designer. You can use the viewers to explore the models. These viewers are also available in SQL Server Management Studio. Each algorithm that you used to build a model in Analysis Services returns a different type of result. Therefore, Analysis Services provides custom viewers for each type of machine learning model. In this lesson you will look at the results from your three models. Each model type is based on a different algorithm and provides different insights into the data.

1. Exploring the Decision Tree Model

Decision tree is a type of supervised learning algorithm (having a pre-defined output variable) that is mostly used in classification problems. It works for both categorical and continuous input and output variables. The output variable can be categorical variable or continuous (numeric) variable. Age and YearlyIncome in the dataset are continuous variables. We can use the decision tree algorithm to predict the customer age or yearly income. In this lesson, the decision tree algorithm predicts which columns influence the decision to purchase a bike (a categorical variable) based upon the remaining columns in the training set.

To explore the model in the Decision Tree tab

1. Select the **Mining Model Viewer** tab in **Data Mining Designer**.
   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.
   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.
   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**.

6. Click on the node for **Age >= 45 and < 52 (the numbers may vary)**. 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 45 and 52 with no car is likely to purchase a bike. Taking it one step further, we find that the likelihood to purchase a bike increases if the customers are actually age 49 to 52.

On the **Decision Tree** tab, you can view decision trees for every predictable attribute in the dataset. In this case, the model predicts only one column, Bike Buyer, so there is only one tree to view. If there were more trees, you could use the **Tree** box to choose another tree. As you view the TM_Decision_Tree model in the Decision Tree viewer, you can see the most important attributes at the left side of the chart. "Most important" means that these attributes have the greatest influence on the outcome. Attributes further down the tree (to the right of the chart) have less of an effect. In this example, age is the single most important factor in predicting bike buying. The model groups customers by age, and then shows the next more important attribute for each age group. For example, in the group of customers aged 34 to 40, the number of cars owned is the strongest predictor after age.

Because you enabled drillthrough when you created the structure and model, you can retrieve detailed information from the model cases and mining structure, including those columns that were not included in the mining model (e.g., emailAddress, FirstName).

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.
2. Right-click a node, and select **Drill Through** then **Model and Structure Columns**.
   The same spreadsheet displays with the structure columns appended to the end.

The **Dependency Network** tab displays the relationships between the attributes that contribute to the predictive ability of the mining model.

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 graph highlights any connected nodes that have an effect on the predictable attribute.
2. Adjust the **All Links** slider to identify the most influential attribute.
   As you drag down the slider, attributes that have only a weak effect on the [Bike Buyer] column are removed from the graph. By adjusting the slider, you can discover that Age and Region are the greatest factors in predicting whether someone is a bike buyer. The Dependency Network viewer reinforces our findings that Age and Region are important factors in predicting bike buying.

2. Exploring the Clustering Model

Clustering is a method of unsupervised learning and is a common technique for statistical data analysis, which is used to see what groups the data points fall into. Common clustering algorithms include K-means clustering, Mean-shift clustering, Density-based spatial clustering, Expectation-maximization clustering, and Agglomerative hierarchical clustering.

Cluster Diagram Tab

The Cluster Diagram tab displays all the clusters that are in a mining model. The lines between the clusters represent "closeness" and are shaded based on how similar the clusters are. The actual color of each cluster represents the frequency of the variable and the state in the cluster.

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.
4. 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.
5. Pause your mouse over the cluster with the darkest shading.
   A tooltip displays the percentage of cases that have the attribute, Bike Buyer = 1.
6. Select the cluster that has the highest density, right-click the cluster, select **Rename Cluster** and type **Bike Buyers High** for later identification. Click **OK**.
7. 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 **OK**.
8. 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.

This document is adapted from Microsoft Tutorial. Credits go to the Microsoft.

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 nonexistent, the clusters are not very similar.

9. 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.

Cluster Profiles Tab

The **Cluster Profiles** tab provides an overall view of the TM_Clustering model. The **Cluster Profiles** tab contains a column for each cluster in the model. The first column lists the attributes that are associated with at least one cluster. The rest of the viewer contains the distribution of the states of an attribute for each cluster. The distribution of a discrete variable is shown as a colored bar with the maximum number of bars displayed in the **Histogram bars** list. Continuous attributes are displayed with a diamond chart, which represents the mean and standard deviation in each cluster.

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**.

Cluster Characteristics Tab

With the **Cluster Characteristics** tab, you can examine in more detail the characteristics that make up a cluster. Instead of comparing the characteristics of all of the clusters (as in the Cluster Profiles tab), you can explore one cluster at a time. For example, if you select **Bike Buyers High** from the **Cluster** list, you can see the characteristics of the customers in this cluster. Though the display is different from the Cluster Profiles viewer, the findings are the same.

Note: Unless you set an initial value for holdoutseed**, results will vary each time you process the model. For more information, see HoldoutSeed Element**

Cluster Discrimination Tab

With the **Cluster Discrimination** tab, you can explore the characteristics that distinguish one cluster from another. After you select two clusters, one from the **Cluster 1** list, and one from the **Cluster 2** list, the viewer calculates the differences between the clusters and displays a list of the attributes that distinguish the clusters most.

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.

3. Exploring the Naïve Bayes Model

Naïve Bayes is a method of supervised learning algorithm for classification. It assumes that the attributes (features) are independent from each other, whereas decision tree algorithm does not have that assumption. The size requirement of dataset is not high. The Microsoft Naive Bayes Viewer provides the following tabs for use in exploring Naive Bayes mining models:

Dependency Network tab

The **Dependency Network** tab works in the same way as the **Dependency Network** tab for the Microsoft Tree Viewer. Each node in the viewer represents an attribute, and the lines between nodes represent relationships. In the viewer, you can see all the attributes that affect the state of the predictable attribute, Bike Buyer.

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**.

This document is adapted from Microsoft Tutorial. Credits go to the Microsoft.

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.

Attribute Profiles tab

The **Attribute Profiles** tab describes how different states of the input attributes affect the outcome of the predictable attribute.

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.
4. 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.
5. Double-click the **Number Cars Owned** cell in the bike buyer (column labeled 1) column.
   The **Mining Legend** displays a more detailed view.

Attribute Characteristics tab

With the **Attribute Characteristics** tab, you can select an attribute and value to see how frequently values for other attributes appear in the selected value cases.

To explore the model in the Attribute Characteristics tab

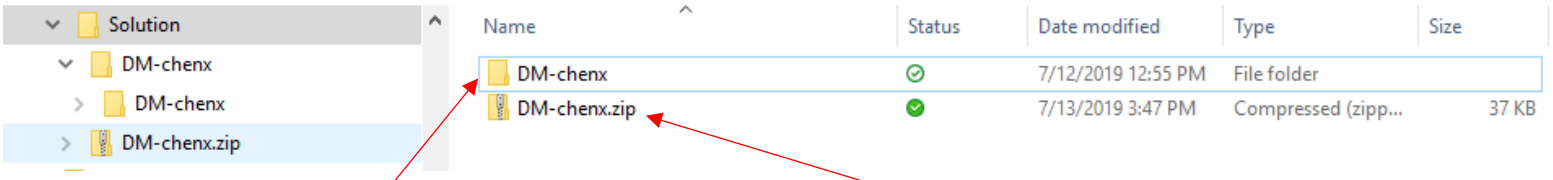
1. In the **Attribute** list, verify that Bike Buyer is selected.
2. Set the **Value** to **1**.
   In the viewer, you will see that customers who have no children at home, short commutes, and live in the North America region are more likely to buy a bike.

Attribute Discrimination tab

With the **Attribute Discrimination** tab, you can investigate the relationship between two discrete values of bike buying and other attribute values. Because the TM_NaiveBayes model has only two states, 1 and 0, you do not have to make any changes to the viewer. In the viewer, you can see that people who do not own cars tend to buy bicycles, and people who own two cars tend not to buy bicycles.

Submission:

1. **Save your project by clicking the Save All button**
2. **Exit from the Visual Studio (SSDT).**
3. **In the Windows Explorer window, locate the project's top folder. As an example, the following Windows Explorer window shows my project DM-chenx stored in a folder called Solution. I expanded the Solution folder in the left navigation pane to show the folder structure of the project. The folder DM-chenx immediately below Solution in the left navigation pane is the top project folder, which is also shown in the right pane of the window.**



4. **Right click the top project folder in the right pane, select Send To -> Compressed (zipped) folder to create a zip file. It should create a zip file with the same name of the top folder, e.g., DM-chenx.zip, right below the top project folder in the right pane of the window as shown in the image above.**
5. **Upload the zip file DM-YourWindowsLogin.zip to Canvas.**
6. **To make sure you have made the right zip file that contains all needed subfolders and files**
   a. **After uploading, download the zip file from Canvas and save it in a location other than the original location**
   b. **Unzip it. In the unzipped folder, double click DM-YourWindowsLoginName.sln (the solution file), e.g., DM-chenx.sln. Then, your project should open in Visual Studio**
   c. **If it cannot open in Visual Studio, it means you didn't zip the top project folder correctly. You want to recreate the zip file following the instructions above.**

This document is adapted from Microsoft Tutorial. Credits go to the Microsoft.