Weekly Exercise 5

NOTE: In this tutorial, we use more algorithms for more business scenarios.

1. **Lesson 1 - creating a new data mining structure for the market basket scenario (association)**
2. **Lesson 2 - building neural network models**
3. **Lesson 3 - building logistic regression models**

We will complete the lessons in the same project created from the previous tutorial.

1. Locate your DM-YourWindowsLoginName.zip, e.g., DM-chenx.zip.
2. Rename the zip file as DM5-YourWindowsLoginName.zip, e.g., DM5-chenx.zip
3. Unzip DM5-YourWindowsLoginName.zip
4. Now you can delete the zip file: DM5-YourWindowsLoginName.zip because after you finish the exercise, you will create a new DM5-YourWindowsLoginName.zip
5. Open your project in Visual Studio by double clicking the DM-YourWindowsLoginName.sln (the solution file for the project) in the project's top folder.
6. IMPORTANT: your username/password for the data source Adventure Works DW2012.ds were not saved. You need to reset the username/password
    a. Double click the data source in Solution Explorer window
    b. On the Data Source Designer window, click the Edit… button
    c. Type your SQL Server username and password again for the corresponding text box. Click the Test Connection button to make sure you can make a connection to the data source. Also, make sure the checkbox for Save my password is checked
    d. Click the Save All button  to save all the changes
    e. Open the Target Mailing and Call Center mining structure by double clicking them in the Solution Explorer window. Click the Mining Model menu, then select Process Mining Structure and All Mining Models… to reprocess all the structure and models
7. Save your project from time to time by clicking the Save All button  to protect the work you have done in case of computer crash

Lesson 1: Creating a new data mining structure for the market basket scenario

The marketing department of Adventure Works Cycles wants to improve the company Web site to promote cross-selling. As part of the site update, they would like the ability to predict products that a customer might want to purchase, based on the other products that are already in the customer's online shopping basket. The marketing department also wants to understand customer purchasing behavior better, so that they can design the Web site so that the items that tend to be purchased together appear together. They have learned that data mining is especially useful for this kind of market basket analysis and have asked you to develop a data mining model. After you complete the tasks in this lesson, you will have a mining model that shows groups of items from historical customer transactions. Additionally, you can use the mining model to predict additional items that a customer may want to purchase. The data source is the same database for the previous tutorial: AdventureWorksDW2012 on misbi.cbe.wwu.edu\multi. Since the data source connection has been created in the previous tutorial, we only need to add the desired data source views for the new lessons.

To create a market basket model, you must use a data source view that supports associative data (a case table and a nested table). This data source view will also be used for the sequence clustering scenario. This data source view is different from others that you may have worked with because it contains a nested table. A nested table is a table that contains multiple rows of information about a single row in the case table. For example, if your model analyzes the purchasing behavior of customers, you would typically use a table that has a unique row for each customer as the case table. However, each customer might make multiple purchases, and you might want to analyze the sequence of purchases, or products that are frequently purchased together. To logically represent these purchases in your model, you add another table to the data source view that lists the purchases for each customer. This nested purchases table is related to the customer table by a many-to-one relationship. The nested table might contain many rows for each customer, each row containing a single product that was purchased, perhaps with additional information about the order that the purchases were made, the price at the time of the order, or any promotions that applied. You can use the information in the nested table as inputs to the model, or as the predictable attribute.

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

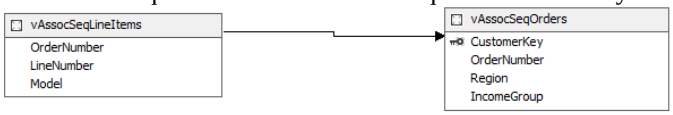
1. To add a new data source view for the market basket analysis

1. In Solution Explorer, right-click **Data Source Views**, and then select **New Data Source View**. The Data Source View Wizard opens.
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 AdventureWorks DW2012 data source that you created in the previous Tutorial. Click **Next**.
4. On the **Select Tables and Views** page, select the following tables, and then click the right arrow to include them in the new data source view:
   - vAssocSeqOrders
   - vAssocSeqLineItems
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 Orders, and then click **Finish**. Open the new data source view by double clicking it in the folder.

2. To create a relationship between tables

1. In Data Source View Designer, position the two tables so that the tables are aligned horizontally, with the vAssocSeqLineItems table on the left side and the vAssocSeqOrders table on the right side.
2. Select the **OrderNumber** column in the vAssocSeqLineItems table.
3. Drag the column to the vAssocSeqOrders table, and put it on the **OrderNumber** column.

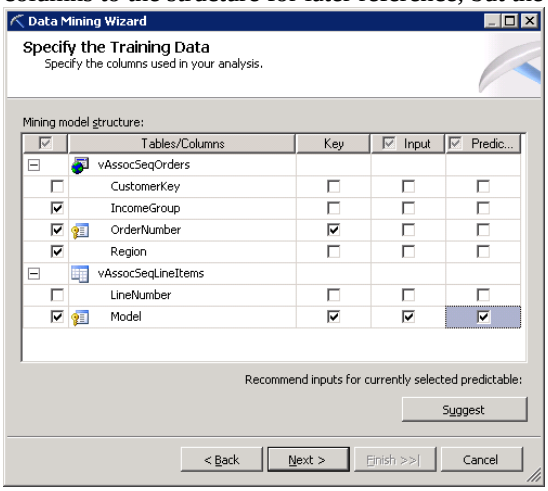
Important: Make sure to drag the **OrderNumber** column from the vAssocSeqLineItems nested table, which represents the many side of the join, to the vAssocSeqOrders case table, which represents the one side of the join. A new *many-to-one relationship* now exists between the vAssocSeqLineItems and vAssocSeqOrders tables. If you have joined the tables correctly, the data source view should appear as follows:



3. To create a market basket structure and model

In this task, you will create a mining structure and a mining model that is based on the Microsoft Association algorithm.

1. In Solution Explorer, right-click **Mining Structures** and select **New Mining Structure** to open the Data Mining Wizard.
2. Click **Next**. On the **Select the Definition Method** page, verify that **From existing relational database or data warehouse** is selected, and then click **Next**.
3. On the **Create the Data Mining Structure** page, under **Which data mining technique do you want to use?**, select **Microsoft Association Rules** from the list, and then click **Next**. The **Select Data Source View** page appears.
4. Select **Orders** under **Available data source views**, and then click **Next**.
5. On the **Specify Table Types** page, in the row for the vAssocSeqLineItems table, select the **Nested** check box, and in the row for the nested table vAssocSeqOrders, select the **Case** check box. Click **Next**.
6. On the **Specify the Training Data** page, clear any boxes that might be checked. Set the key for the case table, vAssocSeqOrders, by selecting the **Key** check box next to OrderNumber.
   Because the purpose of the market basket analysis is to determine which products are included in a single transaction, you do not have to use the **CustomerKey** field.
7. Set the key for the nested table, vAssocSeqLineItems, by selecting the **Key** check box next to Model (product name). The **Input** check box is also automatically selected when you do this. Select the **Predictable** check box for Model as well.
   In a market basket model, you do not care about the sequence of products in the shopping basket, and therefore you should not include **LineNumber** as a key for the nested table. You would use **LineNumber** as a key only in a model where the sequence is important. You will create a model that uses the Microsoft Sequence Clustering algorithm.
8. Select the check box to the left of IncomeGroup and Region, but do not make any other selections. Checking the leftmost column adds the columns to the structure for later reference, but the columns will not be used in the model. Your selections should look like the following:



9. Click **Next**.
10. On the **Specify Columns' Content and Data Type** page, review the selections, which should be as shown in the following table, and then click **Next**.

This document is adapted from [Microsoft Tutorial](). Credits go to the Microsoft.

Columns	Content Type	Data Type
IncomeGroup	Discrete	Text
Order Number	Key	Text
Region	Discrete	Text
vAssocSeqLineItems		
Model	Key	Text

11. On the **Create testing set** page, the default value for the option **Percentage of data for testing** is 30 percent. Change this to **0**. Click **Next**.
    Note: Analysis Services provides different charts for measuring model accuracy. However, some accuracy chart types, such as the lift chart and cross-validation report, are designed for classification and estimation. They are not supported for associative prediction.
12. On the **Completing the Wizard** page, in **Mining structure name**, type Association.
13. In **Mining model name**, type Association.
14. Select the option **Allow drill through**, and then click **Finish**. Data Mining Designer opens to display the Association mining structure that you just created.

4. To process the market basket model

Before you process the association mining model that you created, you must change the default values of two of the parameters: *Support* and *Probability*.

- *Support* defines the percentage of cases in which a rule must exist before it is considered valid. You will specify that a rule must be found in at least 1 percent of cases.
- *Probability* defines how likely an association must be before it is considered valid. You will consider any association with a probability of at least 10 percent.

For more information about the effects of increasing or decreasing support and probability, see Microsoft Association Algorithm Technical Reference. After you have defined the structure and parameters for the **Association** mining model, you will process the model.

To adjust the parameters of the Association model

1. Click the **Mining Models** tab of Data Mining Designer.
2. Right-click the **Association** model column (only model in this window at this moment) in the grid in the designer and select **Set Algorithm Parameters** to open the Algorithm Parameters dialog box.
3. In the **Value** column of the **Algorithm Parameters** dialog box, set the following parameters:
   MINIMUM_PROBABILITY = 0.1
   MINIMUM_SUPPORT = 0.01
4. Click **OK**.

To process the mining model

1. On the **Mining Model** menu, select **Process Mining Structure and All Models.**
2. At the warning asking whether you want to build and deploy the project, click **Yes**.
   The **Process Mining Structure - Association** dialog box opens. Click **Run**.
   The **Process Progress** dialog box opens to display information about model processing. Processing of the new structure and model might take some time.
3. After processing is complete, click **Close** to exit the **Process Progress** dialog box.
4. Click **Close** again to exit the **Process Mining Structure - Association** dialog box.

5. To explore the market basket models

Now that you have built the Association model, you can explore it by using the Microsoft Association Viewer in the **Mining Model Viewer** tab of Data Mining Designer. This tutorial walks you through using the viewer to explore relationships between items. The viewer helps you see at a glance which products tend to appear together, and get a general idea of the emerging patterns. The Microsoft Association Viewer contains three tabs: **Rules**, **Itemsets**, and **Dependency Network**. Because each tab reveals a slightly different view of the data, when you are exploring a model, you will typically switch back and forth between the different panes several times as you pursue insights. For this tutorial, you will start on the **Dependency Network** tab, and then use the **Rules** tab and **Itemsets** tab to deepen your understanding of the relationships revealed in the viewer. You will also use the **Microsoft Generic Content Tree Viewer** to retrieve detailed statistics for individual rules or itemsets.

Dependency Network Tab

With the **Dependency Network** tab, you can investigate the interaction of the different items in the model. Each node in the viewer represents an item, while the lines between them represent rules. By selecting a node, you can see which other nodes predict the selected item, or which items the current item predicts. In some cases, there is a two-way association between items, meaning that they often appear in the same transaction. You can refer to the color legend at the bottom of the tab to determine the direction of the association. A line connecting two items means that these items are likely to appear in a transaction together. In other words, customers are likely to buy these items together. The slider is associated with the probability of the rule. Move the slider up or down to filter out weak associations, meaning rules with low probability. The dependency network graph shows pairwise rules, which can be represented logically as A->B, meaning if Product A is purchased, then Product B is likely. The graph cannot show rules of the type AB->C. If you move the slider to show all rules but still do not see any lines in the graph, it means that there were no pairwise rules that met the criteria of the algorithm parameters. You can also find nodes by name, by typing the first letters of the attribute name. For more information, see Find Node Dialog Box (Mining Model Viewer).

To open the Association mode in the Microsoft Association Rules Viewer

1. In **Solution Explorer**, double-click the Association structure.

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

2. In Data Mining Designer, click the **Mining Model Viewer** tab.
3. Select Association from the list of mining models in the **Mining Model** dropdown list.

To navigate the dependency graph and locate specific nodes

1. In the **Mining Model Viewer** tab, click the **Dependency Network** tab.
2. Click **Zoom In** several times, until you can easily view the labels for each node.
   By default, the graph displays with all nodes visible. In a complex model, there may be many nodes, making each node quite small.
3. Click the + sign in the lower right-hand corner of the viewer and hold down the mouse button to pan around the graph.
4. On the left side of the viewer, drag the slider down, moving it from **All Links** (the default) to the bottom of the slider control.
5. The viewer updates the graph to now show only the strongest association, between the Touring Tire and Touring Tire Tube items.
6. Click the node labeled **Touring Tire Tube = Existing**.
   The graph is updated to highlight only items that are strongly related to this item. Note the direction of the arrow between the two items.
7. On the left side of the viewer, drag the slider up again, moving it from the bottom to around the middle.
   Note the changes in the arrow that connects the two items.
8. Select **Show attribute name only** from the dropdown list at the top of the Dependency Network pane.
   The text labels in the graph are updated to show only the model name.

Itemsets Tab

Next, you will learn more about the rules and itemsets generated by the model for the Touring Tire and Touring Tire Tube products.
The **Itemsets** tab displays three important pieces of information that relate to the itemsets that the Microsoft Association algorithm discovers:

- **Support:** The number of transactions in which the itemset occurs.
- **Size:** The number of items in the itemset.
- **Items:** A list of the items included in each itemset.

Depending on how the algorithm parameters are set, the algorithm might generate many itemsets. Each itemset that is returned in the viewer represents transactions in which the item was sold. By using the controls at the top of the **Itemsets** tab, you can filter the viewer to show only the itemsets that contain a specified minimum support and itemset size.
If you are working with a different mining model and no itemsets are listed, it is because no itemsets met the criteria of the algorithm parameters. In such a scenario, you can change the algorithm parameters to allow itemsets that have lower support.

To filter the itemsets that are shown in the viewer by name

1. Click the **Itemsets** tab of the viewer.
2. In the **Filter Itemset** box, type Touring Tire, and hit the Enter key. The filter returns all items that contain this string.
3. In the **Show** list, select **Show attribute name only**.
4. Select the **Show long name** check box.
   The list of itemsets is updated to show only the itemsets that contain the string Touring Tire. The long name of the itemset includes the name of the table that contains the attribute and value for each item.
5. Clear the **Show long name** check box. The list of itemsets is updated to show only the short name.

The values in the **Support** column indicate the number of transactions for each itemset. A transaction for an itemset means a purchase that included all the items in the itemset. By default, the viewer lists the itemsets in descending order by support. You can click on the column headers to sort by a different column, such as the itemset size or name. If you are interested in learning more about the individual transactions that are included in an itemset, you can drill through from the itemsets to the individual cases. The structure columns in the drillthrough results are the customer's income level and customer ID, which were not used in the model.

To view details for an itemset (This seems not working)

1. In the list of itemsets, click the **Itemset** column heading to sort by name.
2. Locate the item, Touring Tire (with no second item).
3. Right-click the item, Touring Tire, select **Drill Through**, and then select **Model and Structure Columns**.
   The **Drill Through** dialog box displays the individual transactions used as support for this itemset.
4. Expand the nested table, vAssocSeqLineItems, to view the actual list of purchases in the transaction.

To filter itemsets by support or size

1. Clear any text that might be in the **Filter Itemset** box. You cannot use a text filter together with a numeric filter.
2. In the **Minimum support** box, type 1000 (or other higher number than the current number in the box), and then click the background of the viewer.
   The list of itemsets is updated to show only itemsets with support of at least 1000.

Rules Tab

The **Rules** tab displays the following information that is related to the rules that the algorithm finds.

- **Probability:** The *likelihood* of a rule, defined as the probability of the right-hand item given the left-hand side item.
- **Importance:** A measure of the usefulness of a rule. A greater value means a better rule.
  Importance is provided to help you gauge the usefulness of a rule, because probability alone can be misleading. For example, if every transaction contains a water bottle--perhaps the water bottle is added to each customer's cart automatically as part of a promotion--the model would create a rule predicting that water bottle has a probability of 1. Based on probability alone, this rule is very accurate, but it does not provide useful information.
- **Rule:** The definition of the rule. For a market basket model, a rule describes a specific combination of items.

Each rule can be used to predict the presence of an item in a transaction based on the presence of other items. Just like in the **Itemsets** tab, you can filter the rules so that only the most interesting rules are shown. If you are working with a mining model that does not have any rules, you might want to change the algorithm parameters to lower the probability threshold for rules.

To see only rules that include the Mountain-200 bicycle

1. In the **Mining Model Viewer** tab, click the **Rules** tab.

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

2. In the **Filter Rule** box, enter Mountain-200.
   Clear the **Show long name** check box.
3. From the **Show** list, select **Show attribute name only**.
   The viewer will then display only the rules that contain the words "Mountain-200". The probability of the rule tells you how likely it is that when someone buys a Mountain-200 bicycle, that person will also buy the other listed product.

The rules are ordered by probability in descending order, but you can click the column headings to change the sort order. If you are interested in finding out more details about a particular rule, you can use drillthrough to view the supporting cases.

To view cases that support a particular rule

1. In the **Rules** tab, right-click the rule that you want to view.
2. Select **Drill Through**, and then select **Model Columns Only**, or **Model and Structure Columns**.
   The **Drill Through** dialog box provides a summary of the rule at the top of the pane, and a list of all cases that were used as supporting data for the rule.

Lesson 2: Building neural network models for the Operations tasks

The Operations department of Adventure Works is engaged in a project to improve customer satisfaction with their call center. They hired a vendor to manage the call center and to report metrics on call center effectiveness, and have asked you to analyze some preliminary data provided by the vendor. They want to know if there are any interesting findings. In particular, they would like to know if the data suggests any staffing problems with staffing or ways to improve customer satisfaction. The data set is small and covers only a 30-day period in the operation of the call center. The data tracks the number of new and experienced operators in each shift, the number of incoming calls, the number of orders as well as issues that must be resolved, and the average time a customer waits for someone to respond to a call. The data also includes a service quality metric based on *abandon rate*, which is an indicator of customer frustration.

Because you do not have any prior expectations about what the data will show, you decide to use a neural network model to explore possible correlations. Neural network models are often used for exploration because they can analyze complex relationships between many inputs and outputs.

1. To add a new data source view for call center data

In this task, you add a data source view that will be used to access the call center data. The same data will be used to build both the initial neural network model for exploration, and the logistic regression model that you will use to make recommendations. You will also use the Data Source View Designer to add a column for the day of the week. That is because, although the source data tracks call center data by dates, your experience tells you that there are recurring patterns both in terms of call volume and service quality, depending on whether the day is a weekend or a weekday.

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**.
3. On the **Select a Data Source** page, under **Relational data sources**, select the AdventureWorks DW2012 data source you created before.
4. On the **Select Tables and Views** page, select the following table and then click the right arrow to add it to the data source view:
   - **FactCallCenter (dbo)**
   - **DimDate (dbo)**
5. Click **Next**.
6. On the **Completing the Wizard** page, change the name to **CallCenter**, and then click **Finish**. Double click the new data source view to open it.
7. ~~Right-click inside the Data Source View pane, and select **Add/Remove Tables**. Select the table, **DimDate** and click **OK**.~~
   A relationship should be automatically added between the DateKey columns in each table. You will use this relationship to get the column, **EnglishDayNameOfWeek**, from the **DimDate** table and use it in your model.
8. In the Data Source View designer, right-click the table, **FactCallCenter**, and select **New Named Calculation**.
   In the **Create Named Calculation** dialog box, type the following values:

Column name	DayOfWeek
Description	Get day of week from DimDate table
Expression	(SELECT EnglishDayNameOfWeek AS DayOfWeek FROM DimDate where FactCallCenter.DateKey = DimDate.DateKey)

   To verify that the expression creates the data you need, right-click the table **FactCallCenter**, and then select **Explore Data**.
9. Take a minute to review the data that is available, so that you can understand how it is used in data mining:

Column name	**Contains**
FactCallCenterID	An arbitrary key created when the data was imported to the data warehouse. This column identifies unique records and should be used as the case key for the data mining model.
DateKey	The date of the call center operation, expressed as an integer. Integer date keys are often used in data warehouses, but you might want to obtain the date in date/time format if you were going to group by date values. Note that dates are not unique because the vendor provides a separate report for each shift in each day of operation.

This document is adapted from [Microsoft Tutorial](). Credits go to the Microsoft.

Column name	Contains
WageType	Indicates whether the day was a weekday, a weekend, or a holiday. It is possible that there is a difference in quality of customer service on weekends vs. weekdays so you will use this column as an input.
Shift	Indicates the shift for which calls are recorded. This call center divides the working day into four shifts: AM, PM1, PM2, and Midnight. It is possible that the shift influences the quality of customer service so you will use this as an input.
LevelOneOperators	Indicates the number of Level 1 operators on duty. Call center employees start at Level 1 so these employees are less experienced.
LevelTwoOperators	Indicates the number of Level 2 operators on duty. An employee must log a certain number of service hours to qualify as a Level 2 operator.
TotalOperators	The total number of operators present during the shift.
Calls	Number of calls received during the shift.
AutomaticResponses	The number of calls that were handled entirely by automated call processing (Interactive Voice Response, or IVR).
Orders	The number of orders that resulted from calls.
IssuesRaised	The number of issues requiring follow-up that were generated by calls.
AverageTimePerIssue	The average time required to respond to an incoming call.
ServiceGrade	A metric that indicates the general quality of service, measured as the *abandon rate* for the entire shift. The higher the abandon rate, the more likely it is that customers are dissatisfied and that potential orders are being lost.

Note that the data includes four different columns that are based on a single date column: WageType, **DayOfWeek**, Shift, and DateKey. Ordinarily in data mining it is not a good idea to use multiple columns that are derived from the same data, as the values correlate with each other too strongly and can obscure other patterns. However, we will not use DateKey in the model because it contains too many unique values. There is no direct relationship between Shift and **DayOfWeek**, and WageType and **DayOfWeek** are only partly related. If you were worried about collinearity, you could create the structure using all of the available columns, and then ignore different columns in each model and test the effect.

2. To create a neural network structure and model

Because neural networks are extremely flexible and can analyze many combinations of inputs and outputs, you should experiment with several ways of processing the data to get the best results. First, let's create the structure and a default model.

Create the Call Center Structure with a neural network (NN) model

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, verify that the option **Create mining structure with a mining model** is selected.
5. Click the dropdown list for the option **Which data mining technique do you want to use?** Then select **Microsoft Neural Networks**. Because the logistic regression models are based on the neural networks, you can reuse the same structure and add a new mining model. Click **Next**. The **Select Data Source View** page appears.
6. Under **Available data source views**, select CallCenter, and click **Next**.
7. On the **Specify Table Types** page, select the **Case** check box next to the **FactCallCenter** table. Do not select anything for **DimDate**. Click **Next**.
8. On the **Specify the Training Data** page, select **Key** next to the column **FactCallCenterID.**
9. ~~Select the Predictable and **Input** check boxes.~~
10. Select the **Key**, **Input**, and Predict check boxes as shown in the following table:

Tables/Columns	Key/Input/Predict
AutomaticResponses	Input
AverageTimePerIssue	Input/Predict
Calls	Input
Date	Do not use
DateKey	Do not use
DayOfWeek	Input
FactCallCenterID	Key
IssuesRaised	Input
LevelOneOperators	Input/Predict
LevelTwoOperators	Input
Orders	Input/Predict
ServiceGrade	Input/Predict
Shift	Input
TotalOperators	Input
WageType	Input

11. Note that multiple predictable columns have been selected. One of the strengths of the neural network algorithm is that it can analyze all possible combinations of input and output attributes. You wouldn't want to do this for a large data set, as it could exponentially increase processing time.

This document is adapted from [Microsoft Tutorial](). Credits go to the Microsoft.

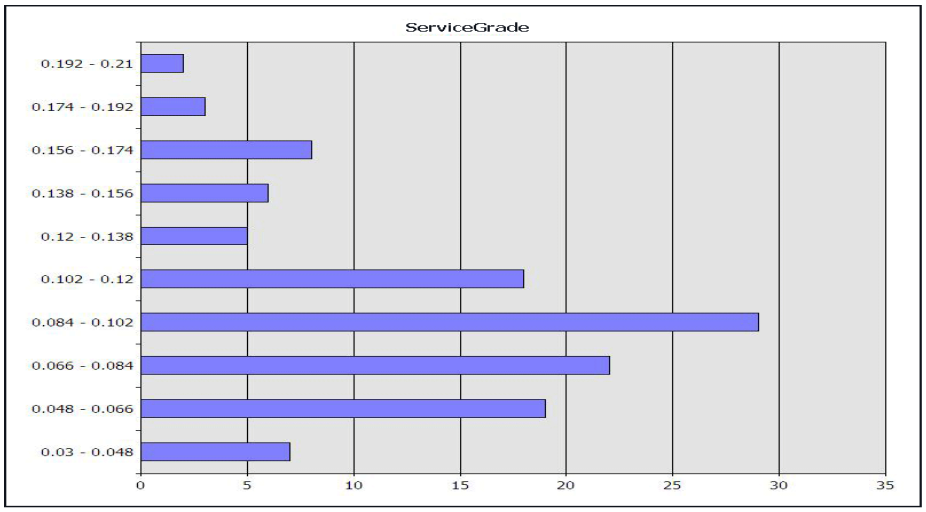
12. On the **Specify Columns' Content and Data Type** page, verify that the grid contains the columns, content types, and data types as shown in the following table, and then click **Next**.

Columns	Content Type	Data Types
AutomaticResponses	Continuous	Long
AverageTimePerIssue	Continuous	Long
Calls	Continuous	Long
DayOfWeek	Discrete	Text
FactCallCenterID	Key	Long
IssuesRaised	Continuous	Long
LevelOneOperators	Continuous	Long
LevelTwoOperators	Continuous	Long
Orders	Continuous	Long
ServiceGrade	Continuous	Double
Shift	Discrete	Text
WageType	Discrete	Text

13. On the **Create testing set** page, clear the text box for the option, **Percentage of data for testing**. Click **Next**.
    1. We can late change the percentage for testing in the Properties window for the mining structure
    2. The property is HoldoutMaxPercent (we can change it to 30, meaning 30%)
14. On the **Completing the Wizard** page, for the **Mining structure name**, type Call Center.
15. For the **Mining model name**, type Call Center Default NN, and then click **Finish**.
    The **Allow drill through** box is disabled because you cannot drill through to data with neural network models.
16. In Solution Explorer, right-click the name of the data mining structure that you just created, and select **Process**. Then run it and close the windows when it is completed.

3. To use discretization to Bin (group) the Target Column

By default, when you create a neural network model that has a numeric predictable attribute, the Microsoft Neural Network algorithm treats the attribute as a continuous number. For example, the ServiceGrade attribute is a number that theoretically ranges from 0.00 (all calls are answered) to 1.00 (all callers hang up). In this data set, the values have the following distribution:



As a result, when you process the model the outputs might be grouped differently than you expect. For example, if you use clustering to identify the best groups of values, the algorithm divides the values in ServiceGrade into ranges such as this one: 0.0748051948 - 0.09716216215. Although this grouping is mathematically accurate, such ranges might not be as meaningful to business users. In this step, to make the result more intuitive, you'll group the numerical values differently, creating copies of the numerical data column.

How Discretization Works

Analysis Services provides a variety of methods for binning or processing numerical data. The following table illustrates the differences between the results when the output attribute ServiceGrade has been processed three different ways:

- Treating it as a continuous number.
- Having the algorithm use clustering to identify the best arrangement of values.
- Specifying that the numbers be binned by the Equal Areas method.

Default model (continuous)

This document is adapted from [Microsoft Tutorial](). Credits go to the Microsoft.

VALUE	SUPPORT
Missing	0
0.09875	120

Binned by clustering

VALUE	SUPPORT
< 0.0748051948	34
0.0748051948 - 0.09716216215	27
0.09716216215 - 0.13297297295	39
0.13297297295 - 0.167499999975	10
>= 0.167499999975	10

Binned by equal areas

VALUE	SUPPORT
< 0.07	26
0.07 - 0.00	22
0.09 - 0.11	36
>= 0.12	36

Note: You can obtain these statistics from the marginal statistics node of the model, after all the data has been processed. For more information about the marginal statistics node, see **Mining Model Content for Neural Network Models (Analysis Services - Data Mining)**.
In this table, the VALUE column shows you how the number for ServiceGrade has been handled. The SUPPORT column shows you how many cases had that value, or that fell in that range.

- **Use continuous numbers (default)**
  If you used the default method, the algorithm would compute outcomes for 120 distinct values, the mean value of which is 0.09875. You can also see the number of missing values.
- **Bin by clustering**
  When you let the Microsoft Clustering algorithm determine the optional grouping of values, the algorithm would group the values for ServiceGrade into five (5) ranges. The number of cases in each range is not evenly distributed, as you can see from the support column.
- **Bin by equal areas**
  When you choose this method, the algorithm forces the values into buckets of equal size, which in turn changes the upper and lower bounds of each range. You can specify the number of buckets, but you want to avoid having two few values in any bucket.

For more information about binning options, see Discretization Methods (Data Mining). Alternatively, rather than using the numeric values, you could add a separate derived column that classifies the service grades into predefined target ranges, such as **Best**(ServiceGrade <= 0.05), **Acceptable** (0.10 > ServiceGrade > 0.05), and **Poor** (ServiceGrade >= 0.10).

To create a Copy of a Column and Change the Discretization Method

You'll make a copy of the mining column that contains the target attribute, ServiceGrade and change the way the numbers are grouped. You can create multiple copies of any column in a mining structure, including the predictable attribute.
For this tutorial, you will use the Equal Areas method of discretization, and specify four buckets. The groupings that result from this method are fairly close to the target values of interest to your business users.

To create a customized copy of a column in the mining structure

1. In Solution Explorer, double-click the mining structure that you just created.
2. In the Mining Structure tab, click the **Add a mining structure column** button.
3. In the **Select column** dialog box, select Service Grade from the list in **Source column**, then click **OK**.
   A new column is added to the list of mining structure columns. By default, the new mining column has the same name as the existing column, with a numerical postfix: for example, Service Grade 1. You can change the name of this column to be more descriptive. You will also specify the discretization method.
4. Right-click Service Grade 1 and select **Properties**.
5. In the **Properties** window, locate the **Name** property, and change the name to **Service Grade Binned** .
6. A dialog box appears asking whether you want to make the same change to the name of all related mining model columns. Click **No**.
7. In the **Properties** window, locate the section **Data Type** (click the Categorized button at the third line of the property window) and expand it if necessary.
8. Change the value of the property Content from Continuous to Discretized.
   The following properties are now available. Change the values of the properties as shown in the following table:

Property	Default value	New value
DiscretizationMethod	Continuous	EqualAreas
DiscretizationBucketCount	No value	4

   Note: The default value of **DiscretizationBucketCount** is actually 0, which means that the algorithm automatically determines the optimum number of buckets. Therefore, if you want to reset the value of this property to its default, type 0.
9. In Data Mining Designer, click the **Mining Models** tab.
   Notice that when you add a copy of a mining structure column, the usage flag for the copy is automatically set to Ignore. Usually, when you add a copy of a column to a mining structure, you would not use the copy for analysis together with the original column, or the algorithm will find a strong correlation between the two columns that might obscure other relationships.

Add a New Mining Model to the Mining Structure

Now that you have created a new grouping for the target attribute, you need to add a new mining model that uses the discretized column. When you are done, the CallCenter mining structure will have two mining models:

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

- The mining model, Call Center Default NN, handles the ServiceGrade values as a continuous range.
- You will create a new mining model, Call Center Binned NN, that uses as its target outcomes the values of the ServiceGrade column, distributed into four buckets of equal size.

To add a mining model based on the new discretized column

1. In Solution Explorer, right-click the mining structure that you just created, and select **Open**.
2. Click the **Mining Models** tab.
3. Click **Create a related mining model**.
4. In the **New Mining Model** dialog box, for **Model name**, type Call Center Binned NN. In the **Algorithm name** dropdown list, select **Microsoft Neural Network**.
5. In the list of columns contained in the new mining model, locate Service Grade, and change the usage from Predict to Ignore.
6. Similarly, locate Service Grade Binned, and change the usage from Ignore to Predict.

Create an Alias for the Target Column

Ordinarily you cannot compare mining models that use different predictable attributes. However, you can create an alias for a mining model column. That is, you can rename the column, Service Grade Binned, within the mining model so that it has the same name as the original column. You can then directly compare these two models in an accuracy chart, even though the data is discretized differently.

To add an alias for a mining structure column in a mining model

1. In the **Mining Models** tab, under **Structure**, select Service Grade Binned.
   Note that the **Properties** window displays the properties of the object, ScalarMiningStructure column.
2. Under the column for the mining model, Call Center Binned NN, click the cell corresponding to the column Service Grade Binned.
   Note that now the **Properties** window displays the properties for the object, MiningModelColumn.
3. Locate the **Name** property, and change the value to Service Grade.
4. Locate the **Description** property and type **Temporary column alias**.
   The **Properties** window should contain the following information:

Property	**Value**
Description	Temporary column alias
ID	Service Grade Binned
Modeling Flags	
Name	Service Grade
SourceColumnID	Service Grade 1
Usage	Predict

5. Click anywhere in the **Mining Model** tab.
   The grid is updated to show the new temporary column alias, Service Grade, beside the column usage. The grid containing the mining structure and two mining models should look like the following:

Structure	**Call Center Default NN**	**Call Center Binned NN**
	Microsoft Neural Network	Microsoft Neural Network
Automatic Responses	Input	Input
Average Time Per Issue	Predict	Predict
Calls	Input	Input
Day Of Week	Input	Input
Fact Call Center ID	Key	Key
Issues Raised	Input	Input
Level One Operators	Input	Input
Level Two Operators	Input	Input
Orders	Input	Input
Service Grade Binned	Ignore	Predict (Service Grade)
Service Grade	Predict	Ignore
Shift	Input	Input
Total Operators	Input	Input
Wage Type	Input	Input

Process All Models

Finally, to ensure that the models you have created can be easily compared, you will set the seed parameter for both the default and binned models. Setting a seed value guarantees that each model starts processing the data from the same point.

Note: If you do not specify a numeric value for the seed parameter, SQL Server Analysis Services will generate a seed based on the name of the model. Because the models always have different names, you must set a seed value to ensure that they process data in the same order.

To specify the seed and process the models

1. In the **Mining Model** tab, right-click the column for the model named Call Center Default NN, and select **Set Algorithm Parameters**.
2. In the row for the HOLDOUT_SEED parameter, click the empty cell under **Value**, and type 1. Click **OK**. Repeat this step for Call Center Binned NN.
   Note: The value that you choose as the seed does not matter, as long as you use the same seed for all related models.
3. In the **Mining Models** menu, select **Process Mining Structure and All Models**. Click **Yes** to deploy the updated data mining project to the server.
4. In the **Process Mining Model** dialog box, click **Run**.

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

5. Click **Close** to close the **Process Progress** dialog box, and then click **Close** again in the **Process Mining Model** dialog box.

Now that you have created the two related mining models, you will explore the data to discover relationships in the data.

4. To explore the call center model

Now that you have built the exploratory model, you can use it to learn more about your data by using the following tools provided in SQL Server Data Tools in Visual Studio.

Microsoft Neural Network Viewer

The viewer has three panes - **Input**, **Output**, and **Variables**.

By using the **Output** pane, you can select different values for the predictable attribute, or dependent variable. If your model contains multiple predictable attributes, you can select the attribute from the **Output Attribute** list. The **Variables** pane compares the two outcomes that you chose in terms of contributing attributes, or variables. The colored bars visually represent how strongly the variable affects the target outcomes. You can also view lift scores for the variables. A lift score is calculated differently depending on which mining model type you are using, but generally tells you the improvement in the model when using this attribute for prediction. The **Input** pane lets you add influencers to the model to try out various what-if scenarios.

Using the Output Pane

In this initial model, you are interested in seeing how various factors affect the grade of service. To do this, you can select Service Grade from the list of output attributes, and then compare different levels of service by selecting ranges from the dropdown lists for **Value 1** and **Value 2**.

To compare lowest and highest service grades

1. For **Value 1**, select the range with the lowest values. For example, the range < 0.07 represents the lowest abandon rates, and therefore the best level of service.
   Note: The exact values in this range may vary depending on how you configured the model.
2. For **Value 2**, select the range with the highest values. For example, the range with the value >=0.12 represents the highest abandon rates, and therefore the worst service grade. In other words, 12% of the customers who phoned during this shift hung up before speaking to a representative.

The contents of the **Variables** pane are updated to compare attributes that contribute to the outcome values. Therefore, the left column shows you the attributes that are associated with the best grade of service, and the right column shows you the attributes associated with the worst grade of service.

Using the Variables Pane

In this model, it appears that Average Time Per Issue is an important factor. This variable indicates the average time that it takes for a call to be answered, regardless of call type.

To view and copy probability and lift scores for an attribute

1. In the **Variables** pane, pause the mouse over the colored bar in the first row.
   This colored bar shows you how strongly Average Time Per Issue contributes toward the service grade. The tooltip shows an overall score, probabilities, and lift scores for each combination of a variable and a target outcome.
2. In the **Variables** pane, right-click any colored bar and select **Copy**.
3. In an Excel worksheet, right-click any cell and select **Paste**.
   The report is pasted as an HTML table, and shows only the scores for each bar.
4. In a different Excel worksheet, right-click any cell and select **Paste Special**.
   The report is pasted as text format, and includes the related statistics described in the next section.

Using the Input Pane

Suppose that you are interested in looking at the effect of a particular factor, such as the shift, or number of operators. You can select a particular variable by using the **Input** pane, and the **Variables** pane is automatically updated to compare the two previously selected groups given the specified variable.

To review the effect on service grade by changing input attributes

1. In the **Input** pane, for **attribute**, select Shift.
2. For **Value**, select **AM**.
   The **Variables** pane updates to show the impact on the model when the shift is **AM**. All other selections remain the same - you are still comparing the lowest and highest service grades.
3. For **Value**, select **PM1**.
   The **Variables** pane updates to show the impact on the model when the shift changes.
4. In the **Input** pane, click the next blank row under **Attribute**, and select Calls. For **Value**, select the range that indicates the greatest number of calls. A new input condition is added to the list. The **Variables** pane updates to show the impact on the model for a particular shift when the call volume is highest.
5. Continue to change the values for Shift and Calls to find any interesting correlations between shift, call volume, and service grade.
   Note: To clear the **Input** pane so that you can use different attributes, click **Refresh viewer content**.

Interpreting the Statistics Provided in the Viewer

Longer waiting times are a strong predictor of a high abandon rate, meaning a poor service grade. This may seem an obvious conclusion; however, the mining model provides you with some additional statistical data to help you interpret these trends.

- **Score**: Value that indicates the overall importance of this variable for discriminating between outcomes. The higher the score, the stronger the effect the variable has on the outcome.
- **Probability of value 1**: Percentage that represents the probability of this value for this outcome.
- **Probability of value 2**: Percentage that represents the probability of this value for this outcome.
- **Lift for Value 1** and **Lift for Value 2**: Scores that represents the impact of using this particular variable for predicting the Value 1 and Value 2 outcomes. The higher the score, the better the variable is at predicting the outcomes.

The following table contains some example values for the top influencers. For example, the **Probability of value 1** is 60.6% and **Probability of value 2** is 8.30%, meaning that when the Average Time Per Issue was in the range of 44-70 minutes, 60.6% of cases were in the shift with the highest service grades (Value 1), and 8.30% of cases were in the shift with the worse service grades (Value 2). From this information, you can draw some conclusions (the actual data may vary). Shorter call response time (the range of 44-70) strongly influences better service grade (the range <0.07). The score (92.35) tells you that this variable is very important.

However, as you look down the list of contributing factors, you see some other factors with effects that are more subtle and more difficult to interpret. For example, shift appears to influence service, but the lift scores and the relative probabilities indicate that shift is not a major factor.

Attribute	**Value**	**Favors < 0.07**	**Favors >= 0.12**
Average Time Per Issue	89.087 - 120.000		Score: 100 Probability of Value1: 4.45 % Probability of Value2: 51.94 % Lift for Value1: 0.19 Lift for Value2: 1.94
Average Time Per Issue	44.000 - 70.597	Score: 92.35 Probability of Value1: 0.06 % Probability of Value2: 8.30 % Lift for Value1: 2.61 Lift for Value2: 0.31	

Lesson 3: Building logistic regression models

In addition to analyzing the factors that might affect call center operations, you were also asked to provide some specific recommendations on how the staff can improve service quality. In this task, you will use the same mining structure that you used to build the exploratory model and add a mining model that will be used for creating predictions. In Analysis Services, a logistic regression model is based on the neural networks algorithm, and therefore provides the same flexibility and power as a neural network model. However, logistic regression is particularly well-suited for predicting binary outcomes. For this scenario, you will use the same mining structure that you used for the neural network model. However, you will customize the new model to target your business questions. You are interested in improving service quality and determining how many experienced operators you need, so you will set up your model to predict those values.

To ensure that all the models based on the call center data are as similar as possible, you will use the same seed value as before. Setting the seed parameter ensures that the model processes the data from the same starting point, and minimizes variations caused by artifacts in the data.

1. To add a new mining model to the call center mining structure

1. In SQL Server Data Tools (SSDT) in Visual Studio, in Solution Explorer, right-click the mining structure, **Call Center**, and select **Open Designer**.
2. In Data Mining Designer, click the **Mining Models** tab.
3. Click **Create a related mining model**.
4. In the **New Mining Model** dialog box, for **Model name**, type Call Center - LR. For **Algorithm name**, select **Microsoft Logistic Regression**. Click **OK**.
   The new mining model is displayed in the **Mining Models** tab.

2. To customize the logistic regression model

1. In the column for the new mining model, Call Center - LR, leave Fact CallCenter ID as the key.
2. Change the value of Service Grade Binned and Level Two Operators to **Predict**.
   These columns will be used both as input and for prediction. In essence, you are creating two separate models on the same data: one that predicts the number of operators, and one that predicts the service grade.
3. Create an alias Service Grade for Service Grade Binned column
   1. Locate and click the cell for the column Service Grade Binned under Call Center-LR
   2. In the Properties window, change the Service Grade Binned to Service Grade for the Name property
4. Change the value of Service Grade to ignore (this column has continuous value, which is not appropriate for a logistic regression model).
5. Change all other columns to **Input**.

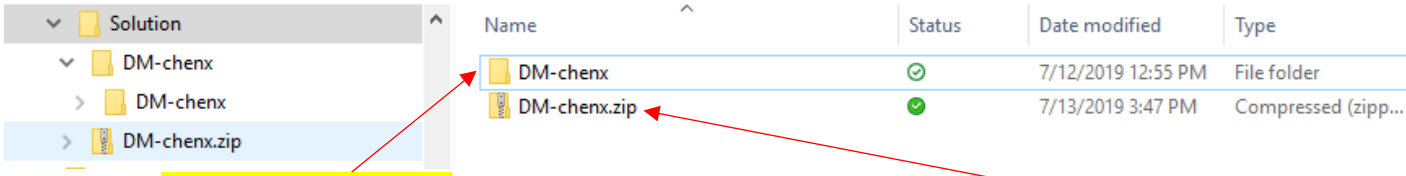
This document is adapted from [Microsoft Tutorial](). Credits go to the Microsoft.

3. To specify the seed and process the models

1. In the **Mining Model** tab, right-click the column for the model named Call Center - LR, and select **Set Algorithm Parameters**.
2. In the row for the HOLDOUT_SEED parameter, click the empty cell under **Value**, and type 1. Click **OK**.
   Note: The value that you choose as the seed does not matter, as long as you use the same seed for all related models.
3. In the **Mining Models** menu, select **Process Mining Structure and All Models**. Click **Yes** to deploy the updated data mining project to the server.
4. In the **Process Mining Model** dialog box, click **Run**.
5. Click **Close** to close the **Process Progress** dialog box, and then click **Close** again in the **Process Mining Model** dialog box.
6. You can use Microsoft Neural Network Viewer to get information of the results from the logistic regression analysis.

Submission:

1. **Save your project by clicking the Save All button** 
2. **Exit from the Visual Studio (SSDT)**
3. **Zip the project's top folder as a new zipped file DM5-YourWindowsLoginName.zip. The following is the example of how to zip the project's top folder**
   a. **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.**



   b. **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.**
4. **Upload the zip file DM5-YourWindowsLoginName.zip to Canvas**
5. **To make sure you have made the right zip file that contains all needed subfolders and files**
   a. **After the 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 DM5-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.