*Hong Kong Baptist University*
***Department of Computer Science***

*COMP 7810/4096 Business Intelligence (2019-20)*

# Data mining in SSAS

## Introduction

**Data Mining** is a process to discover patterns for a large data set. It is an expert system that uses its historical experience (stored in relational databases or cubes) to predict the future.

Imagine that you own a company named Adventureworks. The company sells and manufactures bikes. You want to predict if a customer will buy a bike or not based on the customer information. Data Mining helps you to find the patterns and describe the characteristics of the customers with higher probability to buy the bikes or the lower probability.

Microsoft comes with a nice tool included in SQL Server Analysis Services (SSAS) for creating sophisticated data mining solutions. The tools in Analysis Services help you design, create, and manage data mining models that use either relational or cube data. You can manage client access to the data mining models and create prediction queries from multiple clients.

SSAS contains the features and tools you need to create complex data mining solutions.

- A set of industry-standard data mining algorithms.
- The Data Mining Designer, which you can use to create, manage, and explore data mining models, and then create predictions by using those models.
- The Data Mining Extensions (DMX) language, which you can use to manage mining models and to create complex prediction queries.

Some common data mining algorithms:

- *Clustering*: is a technique to create different groups of people according to their characteristics or patterns. It is a segmentation technique that divides the customers into different groups. We can identify natural groupings of customers.
- *Decision trees*: uses branches to classify the information.

## Learning Outcomes

By finishing this lab session, you should be able to:

- Create a Data Source
- Create a Data View
- Create a Data Mining Project
- Predict information using the Mining Model

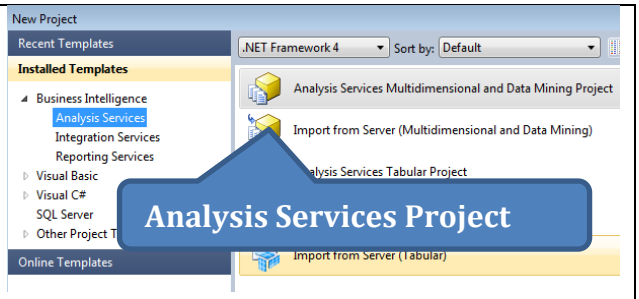
## Tools

- Microsoft SQL Server Management Studio 2012
- Visual Studio 2010 with SQL Server Data Tools (SSDT)

## Part A: Crate data mining task

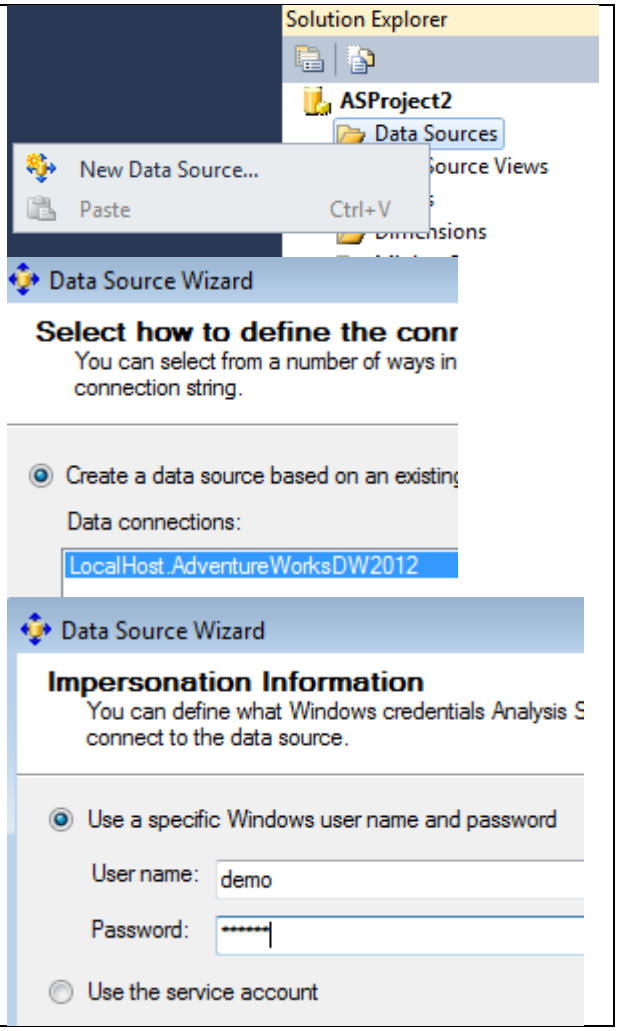
### I. Create a new Analysis Services project

1. Open the **Visual Studio with SSDT.**
2. Select **File → New → Project.**
3. Expand **Business Intelligence → Analysis Services**, and then click **Analysis Services Multidimensional and Data Mining Project**.
4. Change the project name to **ASProject2**. Press **OK**.




### II. Define a new data source and data source view

1. In **Solution Explorer**, **right click Data Sources**, select **New Data Source** then press **Next**.
2. Choose the data connections to connect with AdventureWorksDW2012, press **Next**.
3. In the *Inpersonation Information* dialog box, select **the first option** and **enter your Windows account information**
4. Press **Next**. Then **Finish**.



5. In the **Solution Explorer**, **right click Data Source Views** and select **New Data Source View**. Press **Next** in the Welcome page.

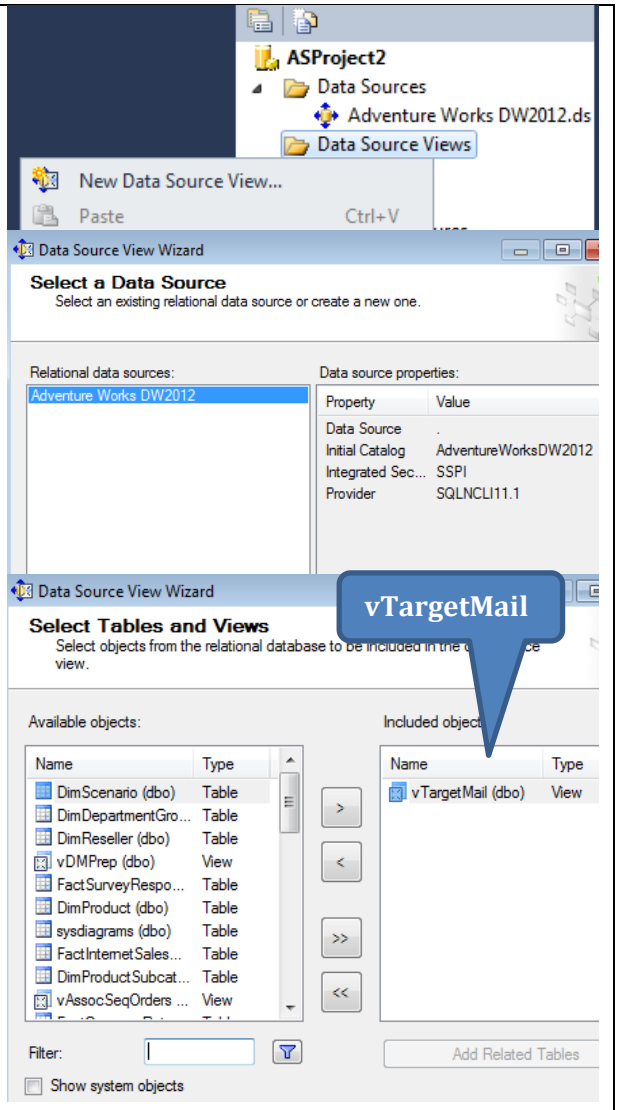
6. In the *Select a Data Source* window, select **Adventure Works DW 2012**. Press **Next**.

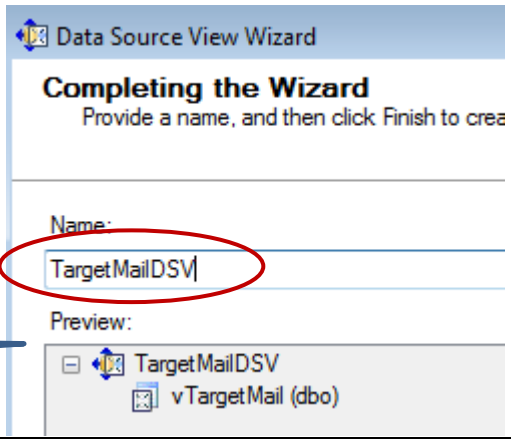
7. In the *Select Tables and Views*, choose **vTargetMail** as an included object.

8. **Rename** the DSV as **TargetMailDSV,** then press **Finish**.

   (You may explore the data inside)

We just created a Data View with the view to give experience to our Data Mining Model. The vTargetMail is a view that contains historical data about the customers. Using that experience, our mining model, will predict the future.



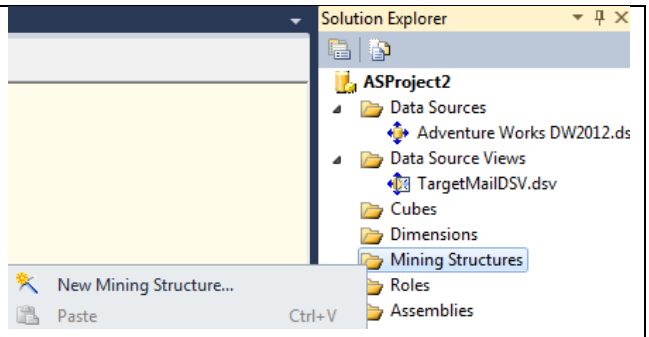




TargetMailDSV

## III. Create New Mining Structures and New Mining Model

1. **Right click** the **Mining Structures** folder and select **New Mining Structure** in the Solution Explorer.

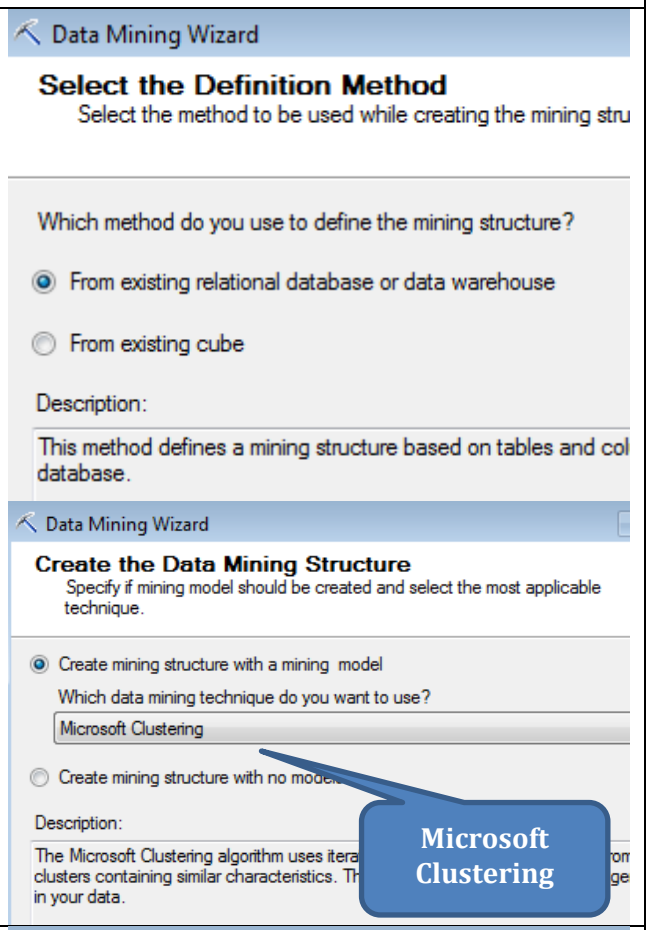
2. It opens up a *Data Mining Wizard*, you can use this wizard to create new mining model for making prediction. Press **Next** to continue.



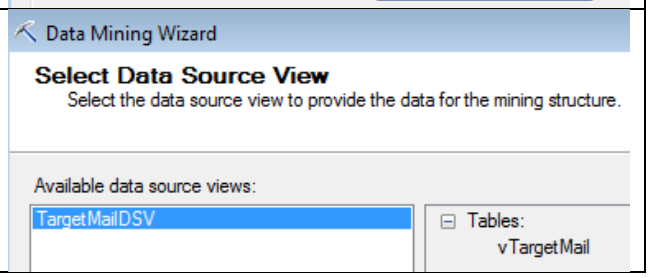

3. In the *select the Definition Method* window, select the option **From existing relational database or data warehouse** to define mining structure. Press **Next** to continue.

4. In the *Create the Data Mining Structure* window, select **Create mining structure with a mining model** → choose **Microsoft Clustering** as the data mining technique. Press **Next**.

   Microsoft Clustering algorithm uses iterative techniques to group records from a dataset into clusters containing similar characteristics. It is used to find general groupings in your data.

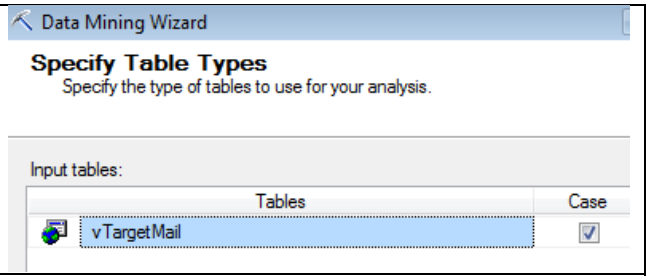



5. In the *Select Data Source View* window, choose **TargetMailDSV.** Press **Next** to continue.



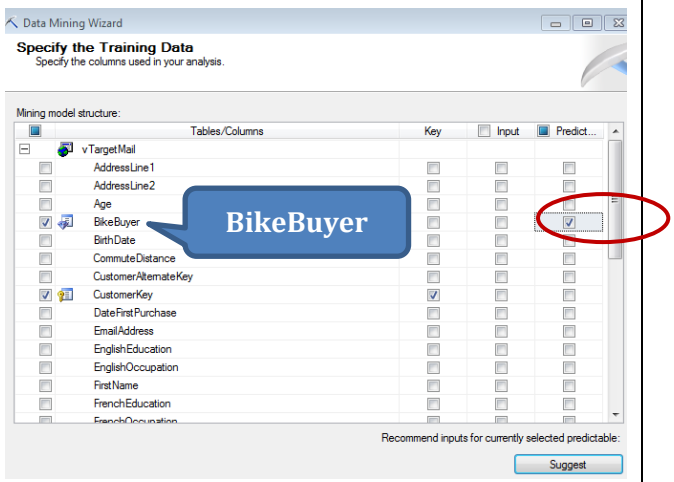
6. Select **vTargetMail** as the input table by default. Press **Next** to continue.

   (This becomes the only one case table)



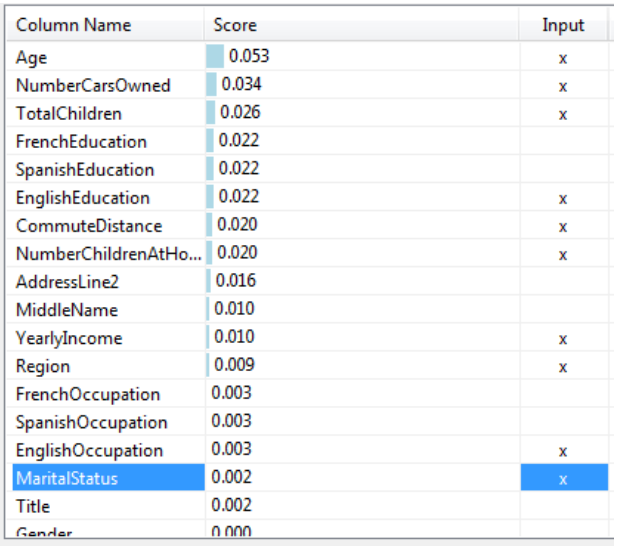

7. Specify the column(s) used in the analysis (which information you want to predict?). **BikeBuyer** is the column in the mining model used for prediction (predictable column). It represents whether the customer will buy the bike or not.

8. Press the button **Suggest** to specify the input columns for the mining model. After sampling, it shows the importance of the attributes within the model with scores greater than or equal to 0.



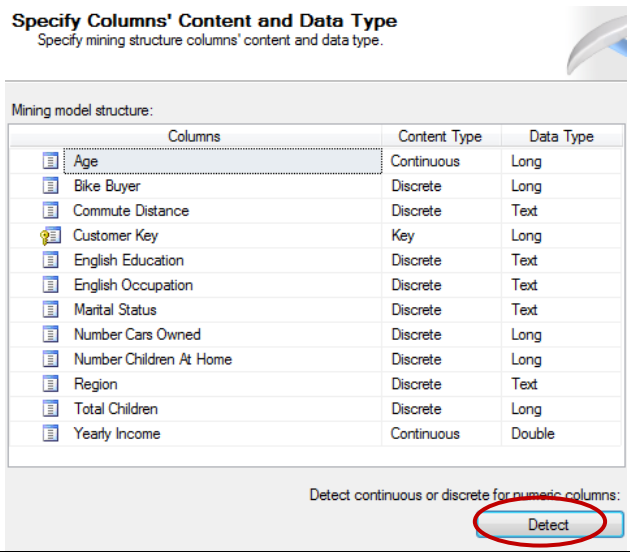

9. Check the following ten attributes as input columns, the press **OK**:
   - Age
   - NumberCarsOwned
   - TotalChildren
   - EnglishEducation
   - CommuteDistance
   - NumberChildrenAtHome
   - YearlyIncome
   - Region
   - EnglishOccupation
   - MaritalStatus

   Press **Next** to continue.



| Column Name | Score | Input |
|---|---|---|
| Age | 0.053 | x |
| NumberCarsOwned | 0.034 | x |
| TotalChildren | 0.026 | x |
| FrenchEducation | 0.022 | |
| SpanishEducation | 0.022 | |
| EnglishEducation | 0.022 | x |
| CommuteDistance | 0.020 | x |
| NumberChildrenAtHo... | 0.020 | x |
| AddressLine2 | 0.016 | |
| MiddleName | 0.010 | |
| YearlyIncome | 0.010 | x |
| Region | 0.009 | x |
| FrenchOccupation | 0.003 | |
| SpanishOccupation | 0.003 | |
| EnglishOccupation | 0.003 | x |
| MaritalStatus | 0.002 | x |
| Title | 0.002 | |
| Gender | 0.000 | |

10. Press **Detect** button to change the data types of the columns. Press **Next** to continue.

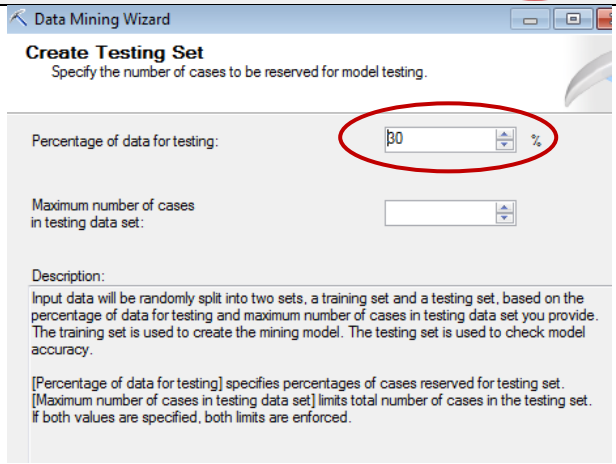



11. Here you can create testing data for testing the accuracy of the DM model, putting **30%** of data for testing means using 70% of data for training the model.

    Note: Input data will be randomly split into two sets, a training set and a testing set, based on the percentage of data for testing

    - *Training set* is used to create the mining model.
    - *Testing set* is used to check model accuracy.

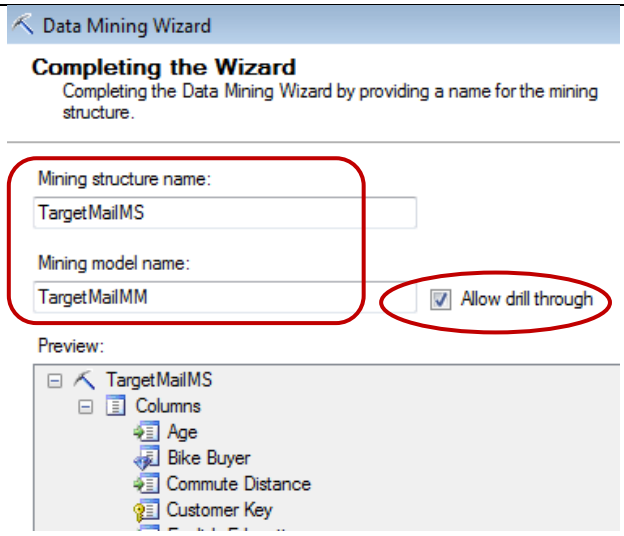
    Press **Next** to continue.



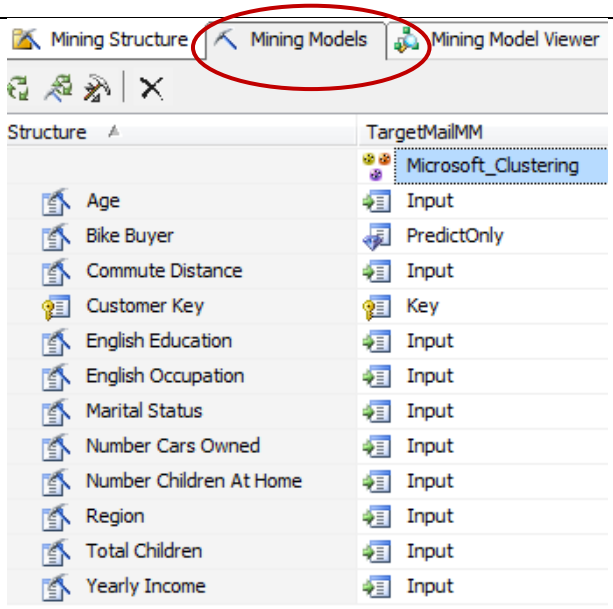

12. Finally provide a name for the structure. Then press **Finish**.
    *Mining structure name*:
    - **TargetMailMS**

    *Mining model name*:
    - **TargetMailMM**

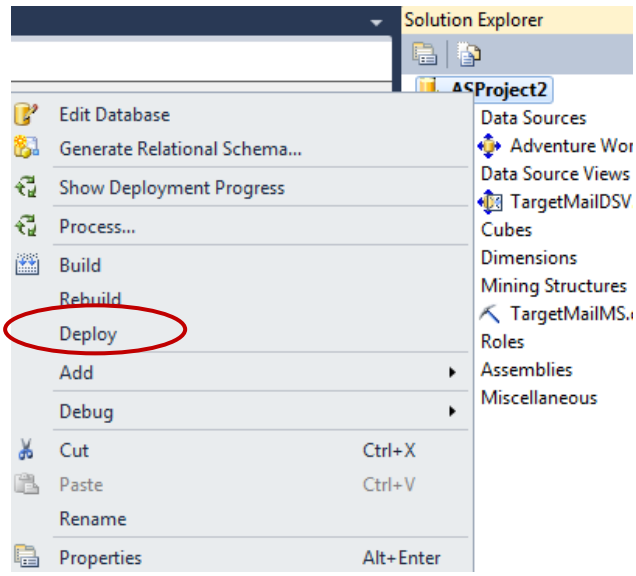
13. Check the box **Allow drill through** and press **Finish**



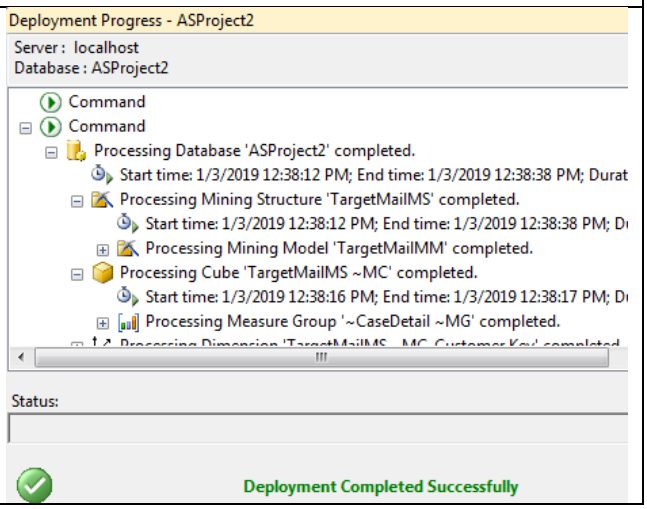
14. Go to **Mining Models** tab, here you can create another data mining model by choosing another data mining algorithm like *Decision Tree*.




15. Go to the **Solution Explorer**, **deploy** the project **ASProject2** to the server. The process will take 70% of data as training data.



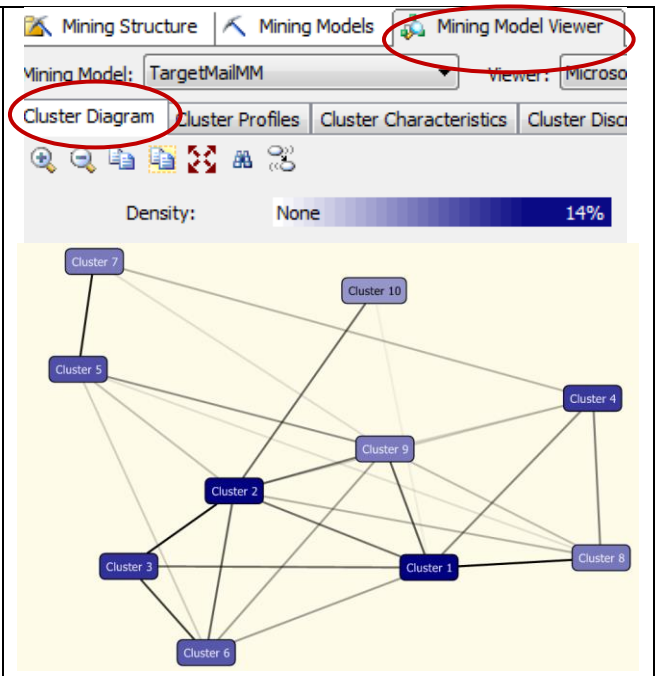

16. You will see the deployment is successful.



17. Select **Mining Model Viewer** tab, the mining model TargetMailMM is created. A **Cluster Diagram** is shown.

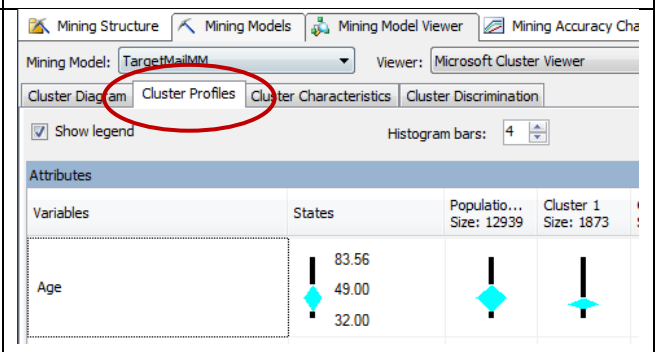
18. Clustering creates different groups for all the customers. The groups are named cluster 1, cluster 2 and so on. The clusters are created based on customers' characteristics. Darker colors means **higher density clusters**.

    (You can change the shading variable to Bike Buyer and state to 1)

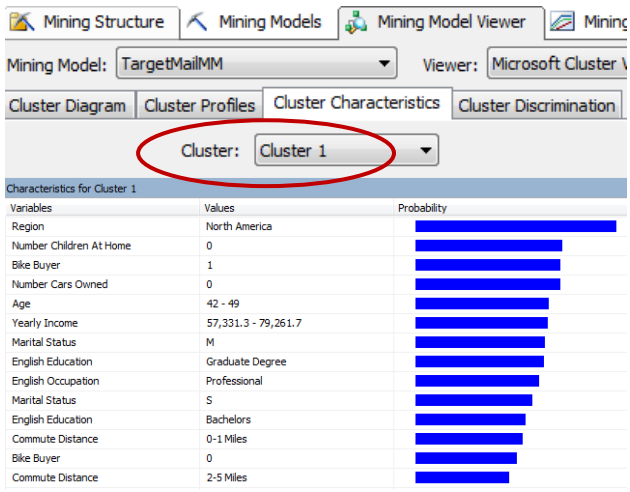



19. Go to **Cluster Profiles** tab, it shows different **variables** and the population for each cluster. The **variables** show the customer's characteristics like the age, yearly income etc

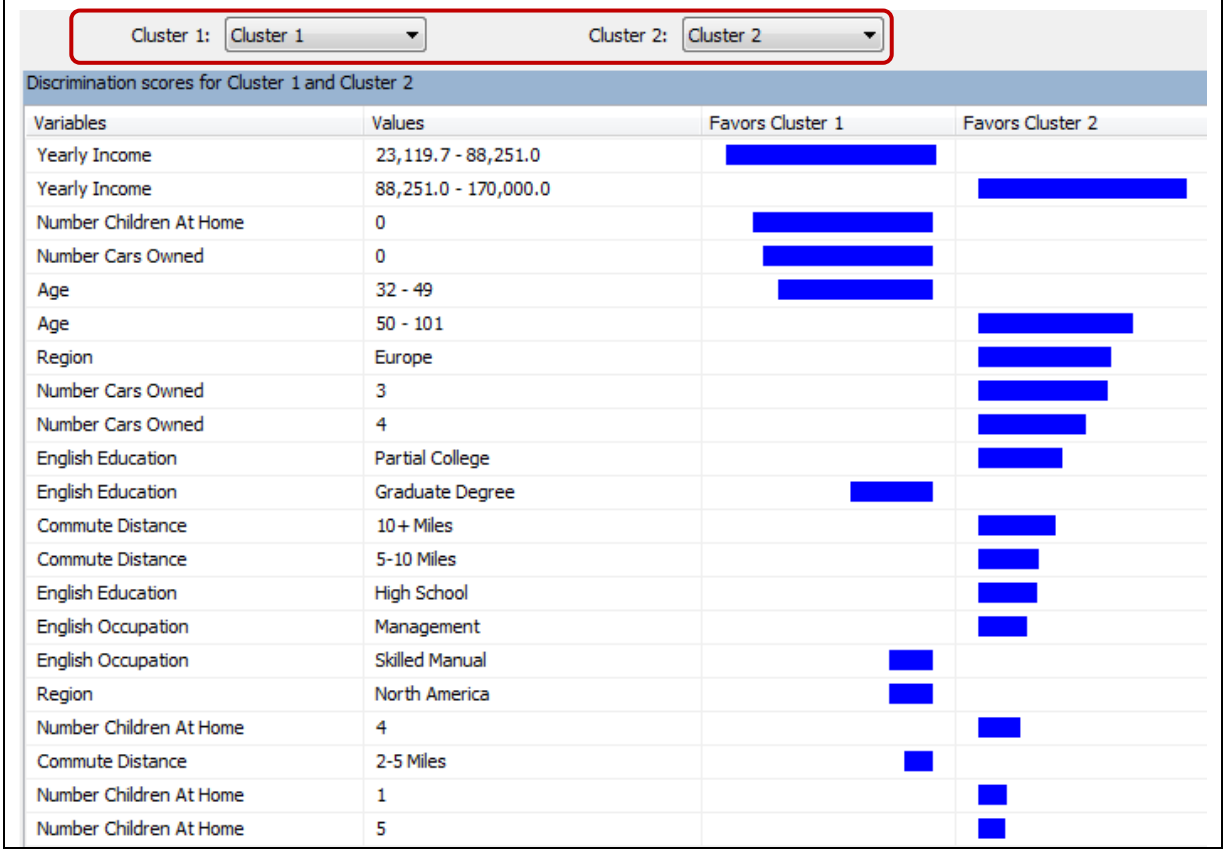
20. The total population is 12939. The *cluster 1* is the *most populated* cluster. 10 clusters are formed according to the population.




21. **Cluster characteristics** tab shows the characteristics per cluster. Select cluster 1, to find out the main characteristics of this cluster.

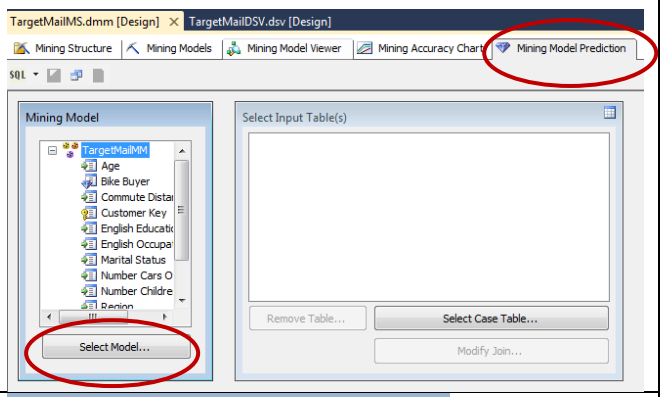
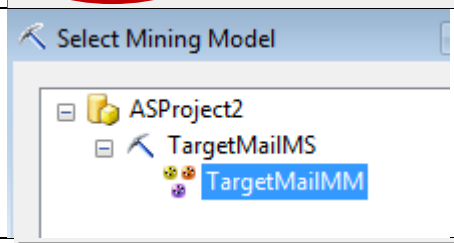
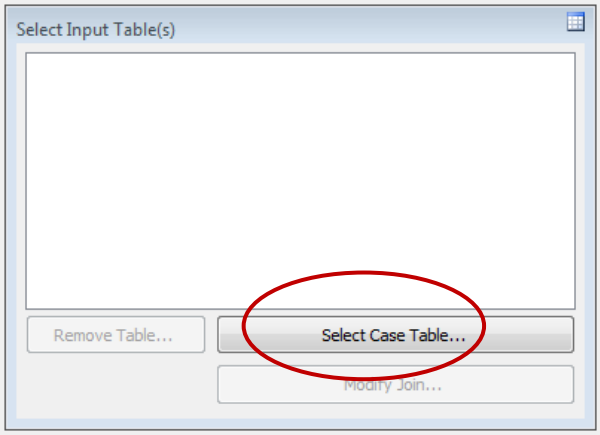
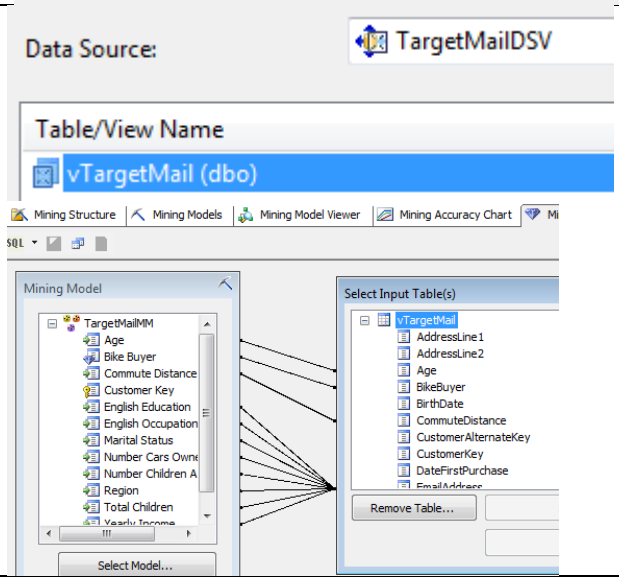



22. **Cluster Discrimination** tab allows you to compare the differences between two clusters.



## IV. Make a Prediction

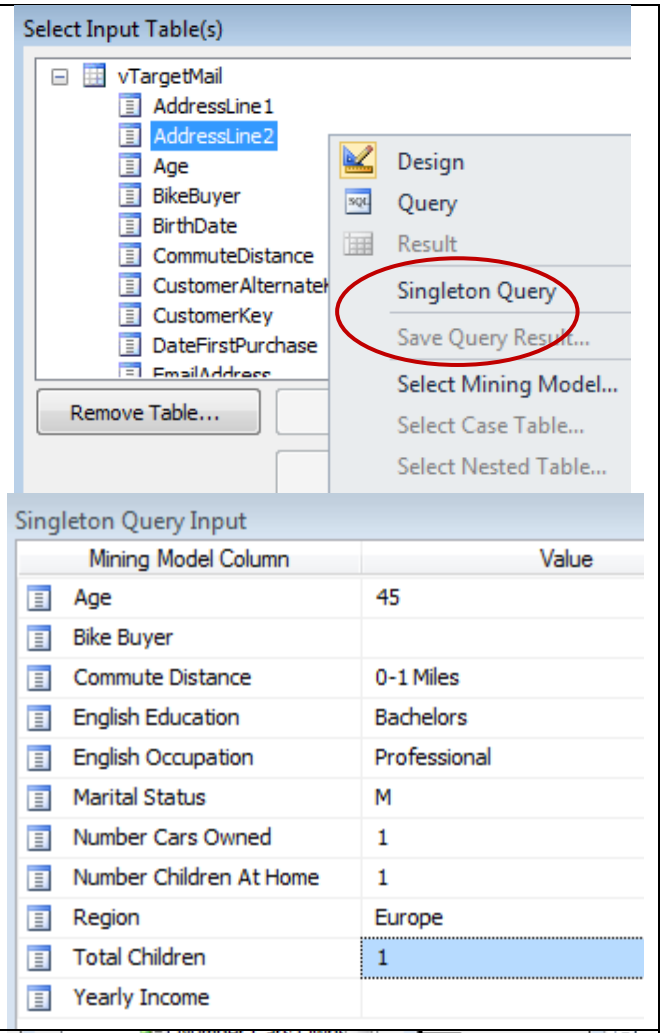
Here are the steps to predict the probability of the customer to buy a bike. 1 represents buy and 0 represents not buy.

| | |
|---|---|
| 1. Click the **Mining Model Prediction** tab, in the **Mining Model** box, press the button **Select Model**. |   |
| 2. In the **Select Mining Model** box, expand **TargetMailMS** and select the model **TargetMailMM**. Then press **OK**. |   |
| 3. In the *Select Input Table(s)* box, click the **Select Case Table** |   |
| 4. Select the **vTargetMail** and press **OK** |   |

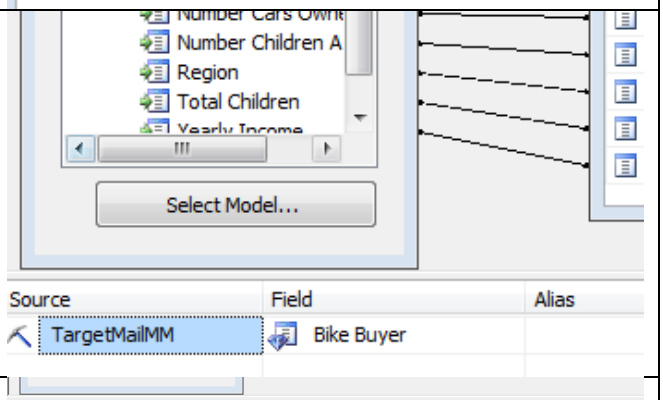
5. **Right click** in the **Select Input Table(s)** and select **Singleton Query**

6. In the **Singleton Query**, specify the following information:
   - Age: 45
   - Commute Distance: 0-1 Miles
   - English Education: Bachelors
   - English Occupation: Professional
   - Marital Status: M
   - Number Cars Owned: 1
   - Number Children At Home: 1
   - Region: Europe
   - Total Children: 1
   - Yearly Income: Missing

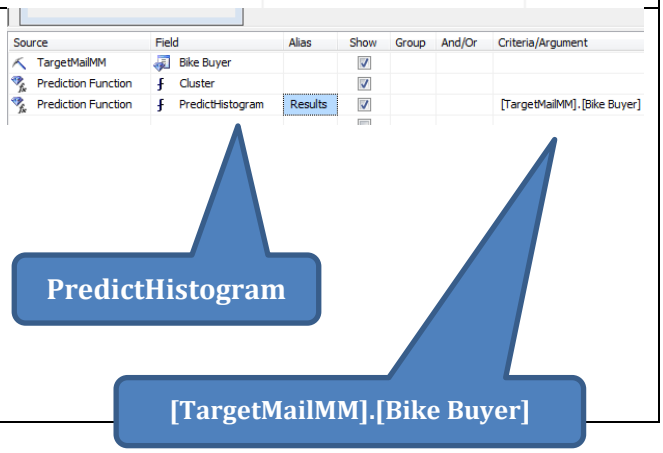
   In this step we are specifying the customer characteristics.




7. In the query grid, for **Source**, select the **TargetMailMM** mining model. (Field is BikeBuyer)



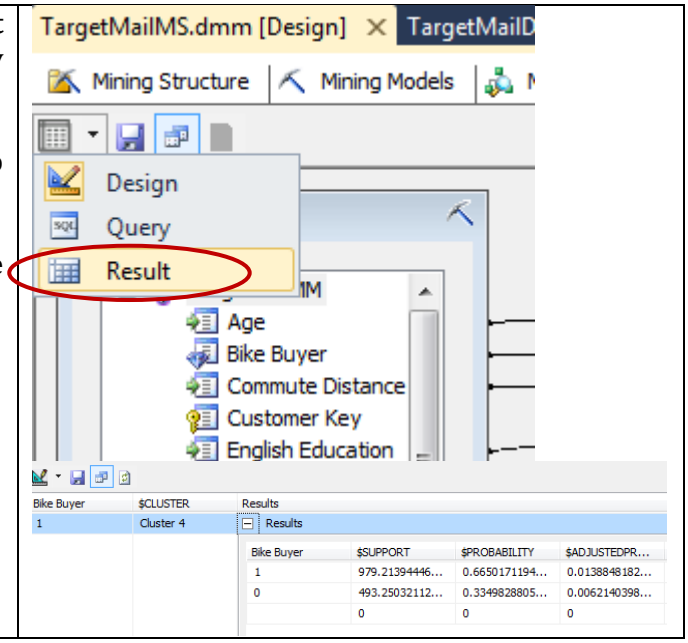

8. In the second row of the **Source** column, select the **Prediction Function**, and add the function **Cluster**.
9. In the second row of the **Source** column, select **Prediction function**, add the function **PredictHistogram,** drag the model column **[Bike Buyer]** into the **Criteria/Argument box**
10. Type **Results** as **alias**



11. Click on the **Switch to query result view** icon and select **Result** to verify the results of the query.

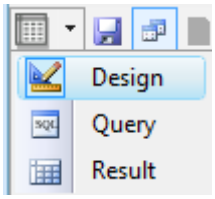
12. The results show the probability to buy a bike is ~66%

13. Select **File → Save All** to save the project **ASProject2**




## V. Exercise

1. Go back to the **Design** View, try to predict another user with the following characteristics will buy a bike or not, submit the result as **online texts** in **buelearning** website.




- Age: 35
- Commute Distance: 1-2 Miles
- English Education: Bachelors
- English Occupation: Clerical
- Marital Status: S
- Number Cars Owned: 1
- Number Children At Home: 0
- Region: North America
- Total Children: 0

The probability that the customer would buy a bike is:

55.6%

The probability that the customer would NOT buy a bike is:

44.3%

## VI. Answer Submission

1. **Zip** your Analysis Services projects (**ASProject2 folder with ASProject2.sln**) that you created in C:\users\demo\documents\visual studio 2010\Projects. The default file name is **ASProject2.zip**

2. Submit the following files to the site **http://buelearning.hkbu.edu.hk/**
   - **ASProject2.zip**
   - **Online texts with probabilities**