## 项目四：数据挖掘项目实施

**1.学时**

**课内学时：4学时**

**课外学时：4学时**

**2.项目目的与要求**

通过实践让学生能灵活运用数据仓库建模过程，联机在线分析过程，并能结合OLAP分析结果进行挖掘算法选择，并建立数据挖掘工程，辅助企业经营决策。

本项目将创建一个目标邮件方案。采用决策树算法、聚类算法、朴素贝叶斯算法三种算法，并通过算法准确度比较，选取准确性最高的算法用于预测潜在客户。

**3.项目环境**

**SQL Server BI**

**4.项目内容**

(1) 在**“数据挖掘设计器”**中，选择**“挖掘模型查看器”**选项卡，默认情况下，设计器将打开添加到结构中的第一个模型（在本例中为 **TM\_Decision\_Tree**），将光标放在标记为**“全部”**的节点上, 将出现哪些信息的工具提示？

(2)单击**“Age >=48 且 < 56”**的节点，能得到哪些客户购买分布情况？

(3) 在“依赖关系网络”选项卡中浏览模型, 单击 **Bike Buyer** 节点以确定它的依赖关系, 调整**“所有链接”**滑块可确定影响最大的属性，向下滑动滑块时，可以发现哪些因素是预测个人自行车购买行为的最主要因素？

(4)使用**“挖掘模型查看器”**选项卡顶部的**“挖掘模型”**列表，切换到 **TM\_Clustering** 模型，使用网络左侧的滑块，可筛选掉强度较低的链接，找出关系最接近的分类有哪些？选择密度最高的分类，右键单击该分类，然后选择**“重命名分类”**并键入 **Bike Buyers High** 以用作日后标识，与之关联的设置为**Bike Buyers Low**。单击“确定”。

(5) 在“分类对比”选项卡中浏览模型，在**“分类 1”**框中，选择 **Bike Buyers High，**在**“分类 2”**框中，选择 **Bike Buyers Low，Bike Buyers Low** 和 **Bike Buyers High** 分类中的客户之间的其他一些显著差异包括哪些内容？

(6) 使用**“挖掘模型查看器”**选项卡顶部的**“挖掘模型”**列表切换到 **TM\_NaiveBayes** 模型，向下滑动滑块时，将只保留对 [Bike Buyer] 列影响最大的属性。通过调整滑块，可以发现影响最大的几个属性为？

**5.项目步骤**

(1)新建数据源视图，选择表**ProspectiveBuyer (dbo)** - 预期自行车购买者的表和表**vTargetMail (dbo)** - 有关以前的自行车购买者的历史数据的视图。

(2) 创建用于目标邮件方案的挖掘结构，在解决方案资源管理器中，右键单击**“挖掘结构”**并选择**“新建挖掘结构”**启动数据挖掘向导，在**“创建数据挖掘结构”**页的**“您要使用何种数据挖掘技术？”**下，选择**“Microsoft 决策树”**。

(3) 在**“指定表类型”**页上，选中 vTargetMail 的**“事例”**列中的复选框以将其用作事例表，然后单击**“下一步”**。稍后您将使用 ProspectiveBuyer 表进行测试，不过现在可以忽略它。

(4) 在**“指定定型数据”**页上，您将为模型至少标识一个可预测列、一个键列以及一个输入列。选中 **BikeBuyer** 行中的**“可预测”**列中的复选框。

(5) 确认在 **CustomerKey** 行中已选中**“键”**列中的复选框。

(6) 选中以下行中**“输入”**列中的复选框。可通过下面的方法来同时选中多个列：突出显示一系列单元格，然后在按住 Ctrl 的同时选中一个复选框。

* + Age
  + CommuteDistance
  + EnglishEducation
  + EnglishOccupation
  + Gender
  + HouseOwnerFlag
  + MaritalStatus
  + NumberCarsOwned
  + NumberChildrenAtHome
  + Region
  + TotalChildren
  + YearlyIncome

(7) 在该页的最左侧的列中，选中以下行中的复选框。

* + AddressLine1
  + AddressLine2
  + DateFirstPurchase
  + EmailAddress
  + FirstName
  + LastName

(8) 检查和修改每列的内容类型和数据类型。

|  |  |  |
| --- | --- | --- |
| 列 | 内容类型 | 数据类型 |
| Address Line1 | Discrete | Text |
| Address Line2 | Discrete | Text |
| Age | Continuous | Long |
| Bike Buyer | Discrete | Long |
| Commute Distance | Discrete | Text |
| CustomerKey | Key | Long |
| DateFirstPurchase | Continuous | Date |
| Email Address | Discrete | Text |
| English Education | Discrete | Text |
| English Occupation | Discrete | Text |
| FirstName | Discrete | Text |
| Gender | 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 |

(9) 在**“创建测试集”**页上，将**“测试数据百分比”**保留其默认值：**30;** 对于**“测试数据集中的最大事例数”**，请键入 **1000**。

(10)命名模型和结构并指定钻取，选中**“允许钻取”**复选框。

(11) 在已经创建的包含一个基于 Microsoft 决策树算法的挖掘模型的挖掘结构中创建两个附加模型，然后处理和部署这些模型，使用 [Microsoft 聚类分析算法](http://msdn.microsoft.com/zh-cn/library/ms174879.aspx)和 [Microsoft Naive Bayes 算法](http://msdn.microsoft.com/zh-cn/library/ms174806.aspx)。

(12) 切换到 Business Intelligence Development Studio 中数据挖掘设计器的**“挖掘模型”**选项卡，右键单击**“结构”**列，选择**“新建挖掘模型”，** 在**“新建挖掘模型”**对话框中的**“模型名称”**中，键入 **TM\_Clustering，**在**“算法名称”**中，选择**“Microsoft 聚类分析”。**

(13) 在数据挖掘设计器的**“挖掘模型”**选项卡中，右键单击**“结构”**列，并选择**“新建挖掘模型”，** 在**“新建挖掘模型”**对话框中的**“模型名称”**下，键入 **TM\_NaiveBayes，**在**“算法名称”**中，选择 **Microsoft Naive Bayes**，再单击**“确定”，** 此时将显示一条消息，说明 Microsoft Naive Bayes 算法不支持 **Age** 和 **Yearly Income** 列，这些都是连续列，单击**“是”**，以确认此消息并继续下面的操作。

(14) 在 Business Intelligence Development Studio 的数据挖掘设计器中，单击**“挖掘结构”**选项卡或**“挖掘模型”**选项卡，**Targeted Mailing MiningStructure** 显示在**“属性”**窗格中，按 **F4** 可以打开**“属性”**窗格，确保 **CacheMode** 已设置为 **KeepTrainingCases，**为 **HoldoutSeed** 输入 **12。**

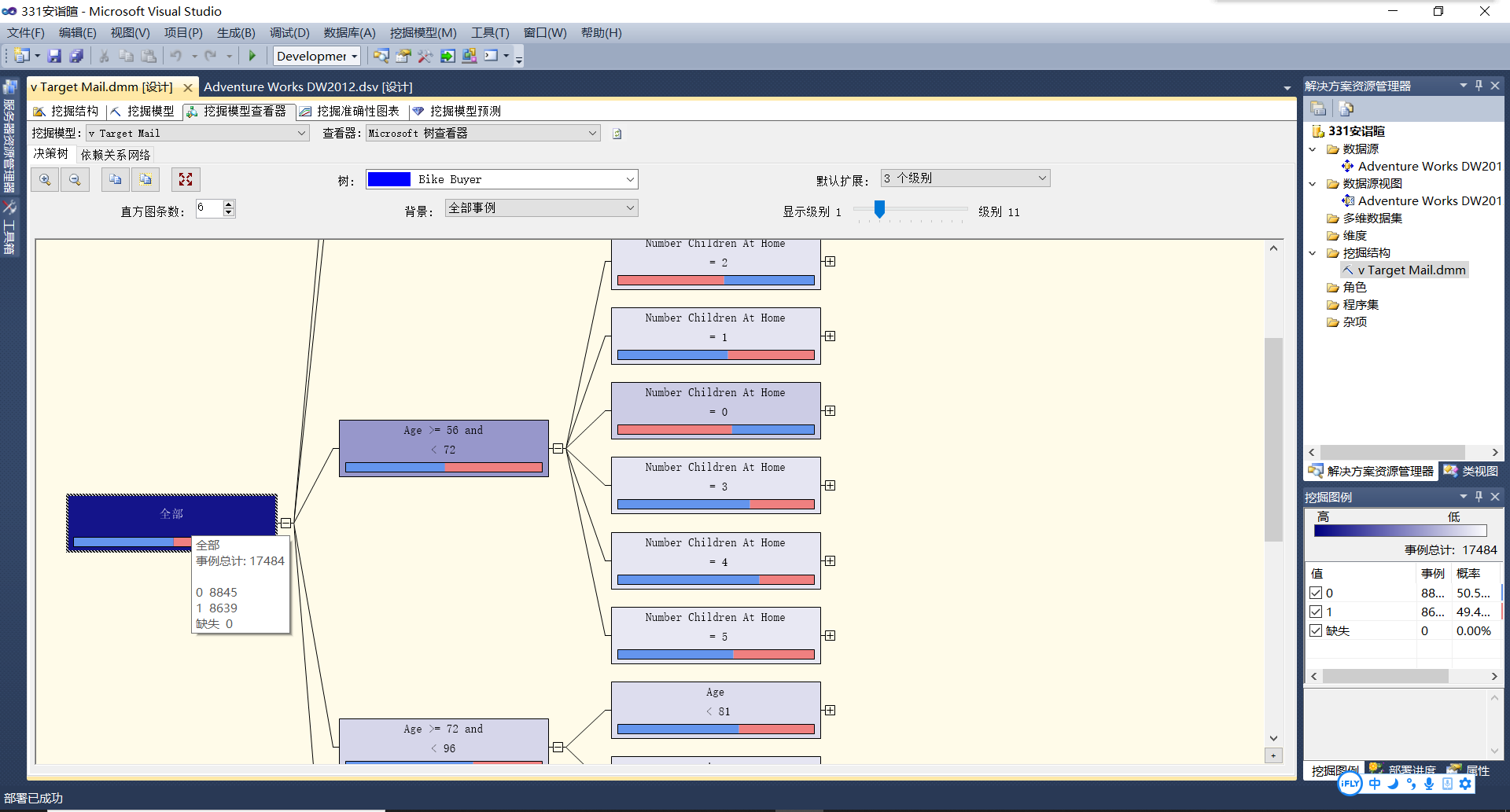
(15) 部署并处理模型。

**6.成果物**

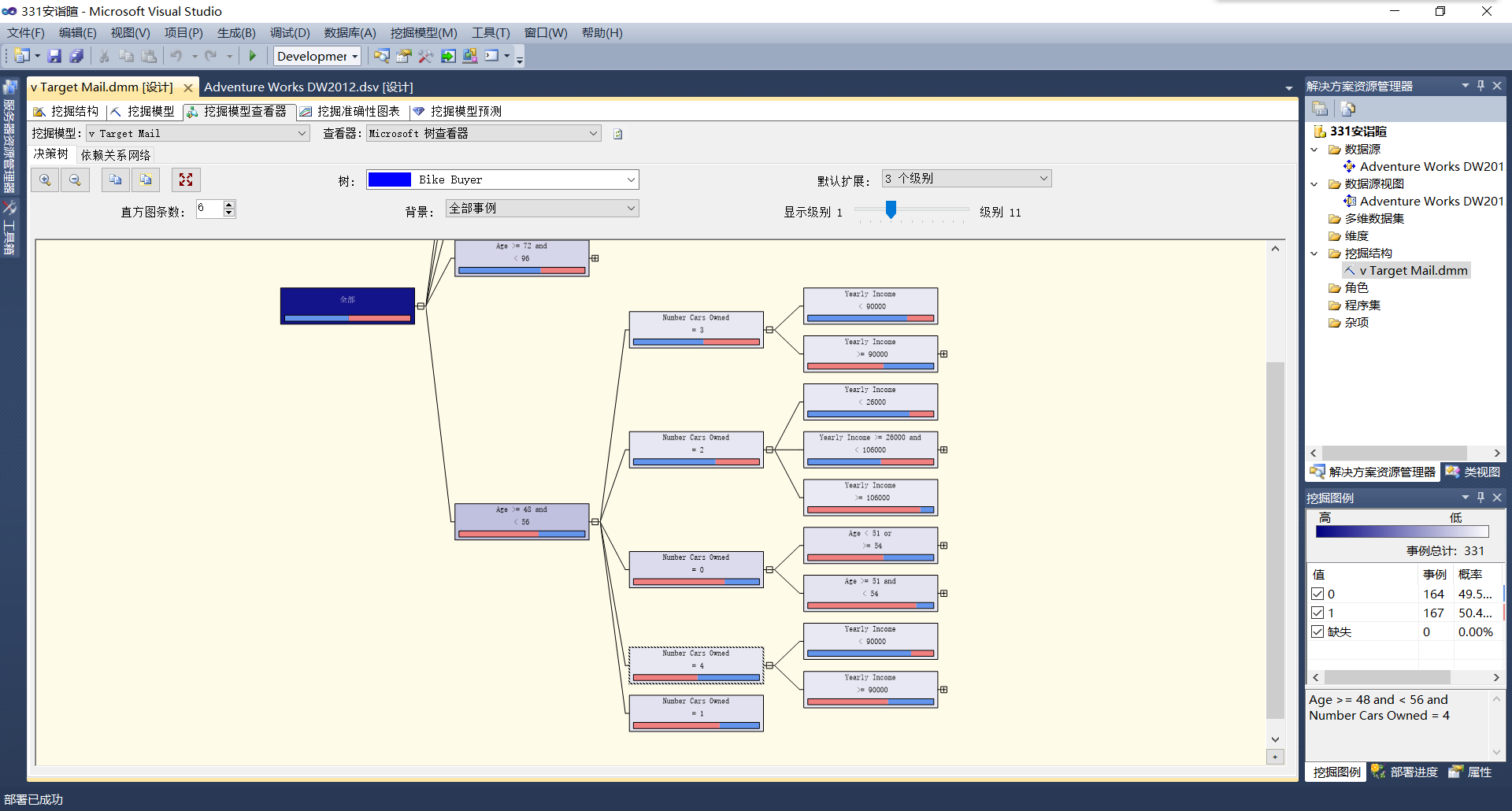
实验成果展示

(1) 在**“数据挖掘设计器”**中，选择**“挖掘模型查看器”**选项卡，默认情况下，设计器将打开添加到结构中的第一个模型（在本例中为 **TM\_Decision\_Tree**），将光标放在标记为**“全部”**的节点上, 将出现哪些信息的工具提示？

**将出现的信息工具提示有“事例总计”、“缺失”。**

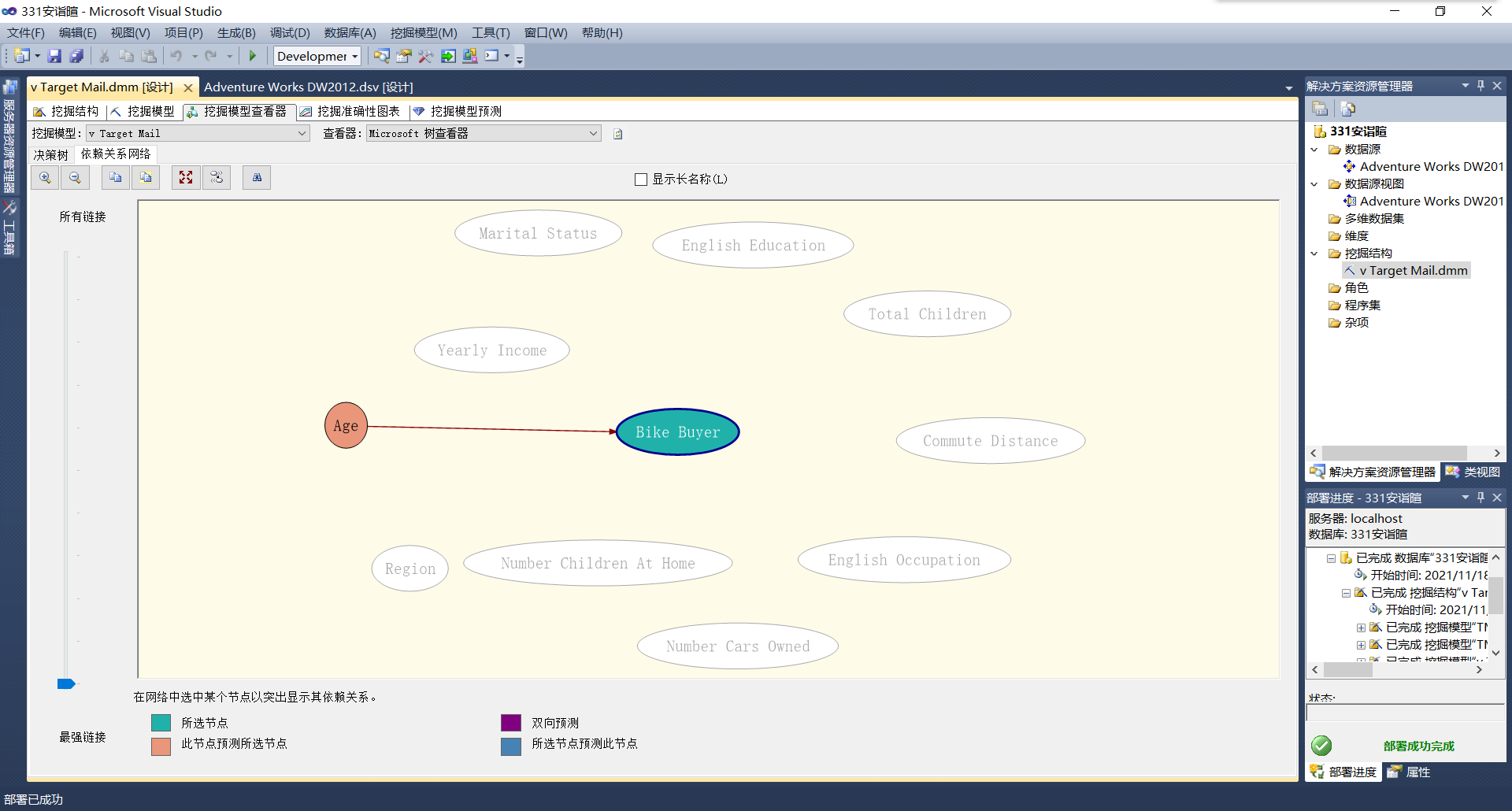


(2)单击**“Age >=48 且 < 56”**的节点，能得到哪些客户购买分布情况？



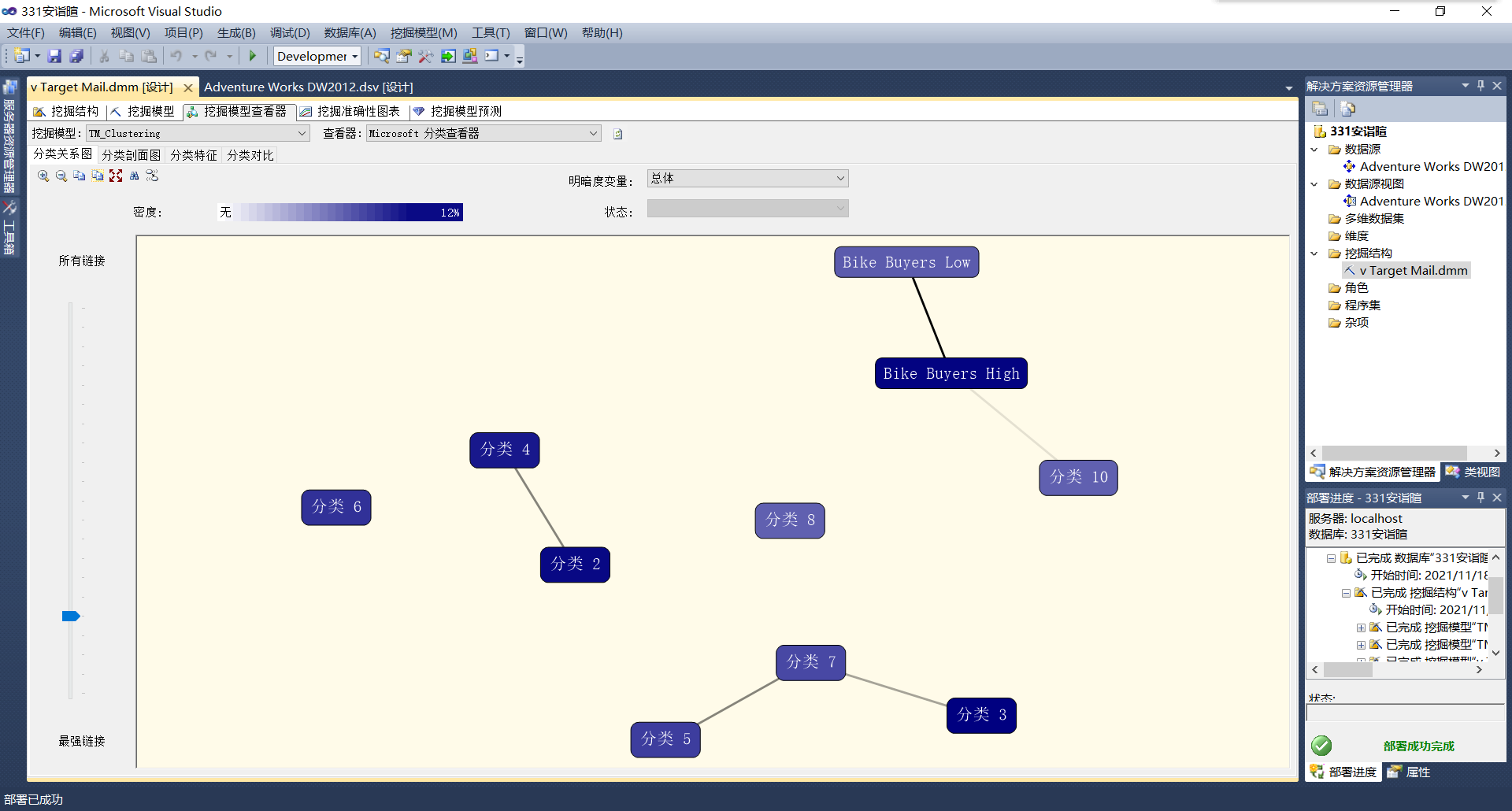
(3) 在“依赖关系网络”选项卡中浏览模型, 单击 **Bike Buyer** 节点以确定它的依赖关系, 调整**“所有链接”**滑块可确定影响最大的属性，向下滑动滑块时，可以发现哪些因素是预测个人自行车购买行为的最主要因素？

**Age是预测个人自行车购买行为的最主要因素。**



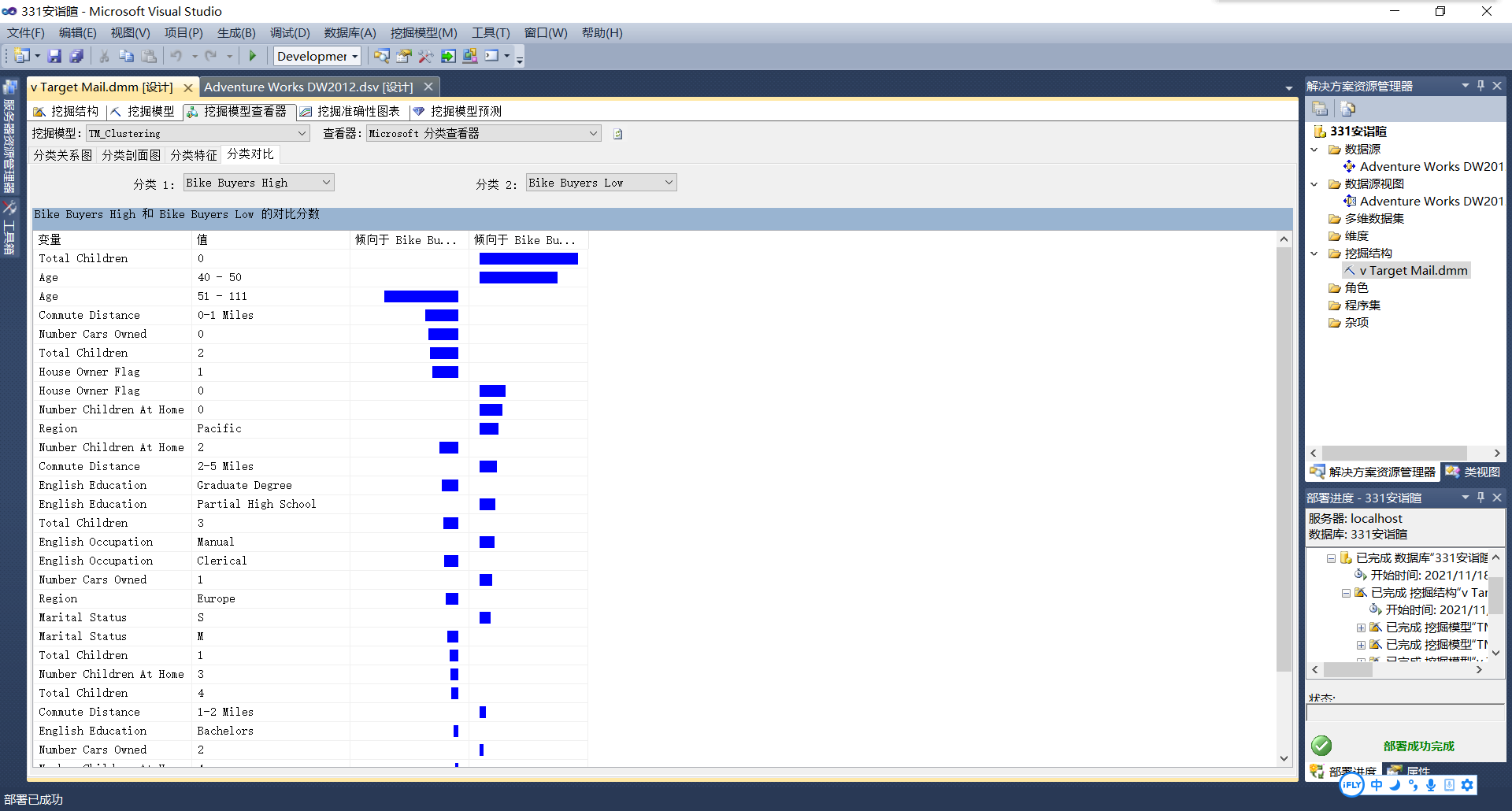
(4)使用**“挖掘模型查看器”**选项卡顶部的**“挖掘模型”**列表，切换到 **TM\_Clustering** 模型，使用网络左侧的滑块，可筛选掉强度较低的链接，找出关系最接近的分类有哪些？选择密度最高的分类，右键单击该分类，然后选择**“重命名分类”**并键入 **Bike Buyers High** 以用作日后标识，与之关联的设置为**Bike Buyers Low**。单击“确定”。

**关系最接近的分类为分类1和分类9。**



(5) 在“分类对比”选项卡中浏览模型，在**“分类 1”**框中，选择 **Bike Buyers High，**在**“分类 2”**框中，选择 **Bike Buyers Low，Bike Buyers Low** 和 **Bike Buyers High** 分类中的客户之间的其他一些显著差异包括哪些内容？

**显著差异包括Total Children、Age、Commute Distance。**



(6) 使用**“挖掘模型查看器”**选项卡顶部的**“挖掘模型”**列表切换到 **TM\_NaiveBayes** 模型，向下滑动滑块时，将只保留对 [Bike Buyer] 列影响最大的属性。通过调整滑块，可以发现影响最大的几个属性为？

**影响最大的属性为Number Cars Owned。**

