

您的潜力，我们的动力

Microsoft
微软(中国)有限公司

SQL2005 数据挖掘算法详解

Lesson 1

杨大川

CTO

北京迈思奇科技有限公司



MSDN Webcasts

讲师简介

- 杨大川 - 迈思奇科技有限公司CTO
 - 微软MVP.2004, 2005 (最有价值专家)
 - 曾任美国硅谷Annuncio公司首席工程师
 - 招商迪辰产品研发部总经理
 - 现兼任中科院客座教授
- Minesage :迈思奇科技有限公司
 - 微软数据分析/挖掘领域合作伙伴
 - 面向企业客户提供完整的数据分析与挖掘解决方案
 - 提供专业、高端的BI培训
 - www.minesage.com



您的潜力. 我们的动力

Microsoft
微软(中国)有限公司

收听本次课程需具备的条件

- 本讲座难度属于中级
- 面向技术人员

本次课程内容包括

- SQL2005数据挖掘概述
- 贝叶斯 (Naive Bayes)
- 决策树 (Decision Trees)
 - 线性回归 (Linear Regression)
- 神经网络 (Neural Networks)
 - 逻辑回归 (Logistic Regression)
- 比较挖掘的准确度

什么是数据挖掘?

您的潜
Micro
微软(中国)



数据挖掘(**Data Mining**), 又称信息发掘 (**Knowledge Discovery**), 是用自动或半自动化的方法在数据中找到潜在的, 有价值的信息和规则.

数据挖掘技术来源于数据库, 统计和人工智能.



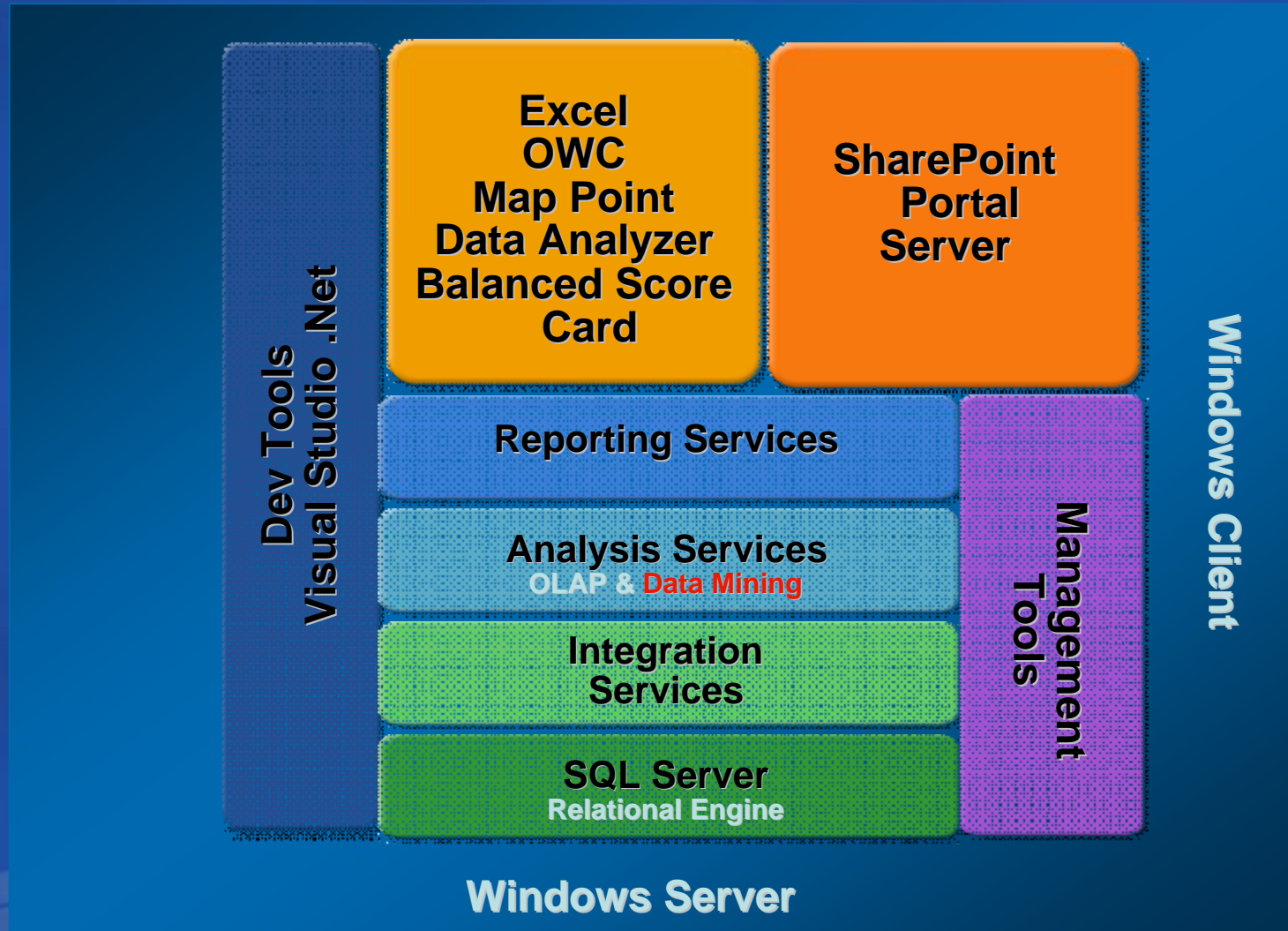
Microsoft®
微软(中国)有限公司

```
graph LR; HD[(Historical Dataset)] --> ETL[Data Transform ETL]; C1[Cube] --> MM[Mining Models]; ND[(New Dataset)] --> MM; MM --> MB[Model Browsing]; MM --> R[Reporting]; MM --> P{Prediction}; R --> C1; P --> C1; P --> LOB[LOB Application];
```

您的潜力, 我们的动力

数据挖掘与 Microsoft 商务智能

Microsoft
微软(中国)有限公司



API: DMX

CREATE MINING MODEL CreditRisk

(CustID LONG KEY,
Gender TEXT DISCRETE,
Income LONG CONTINUOUS,
Profession TEXT DISCRETE,
Risk TEXT DISCRETE PREDICT)

USING Microsoft_Decision_Trees

INSERT INTO CreditRisk

(CustId, Gender, Income, Profession,
Risk)

Select

CustomerID, Gender, Income,
Profession, Risk

From Customers

Select NewCustomers.CustomerID, CreditRisk.Risk,
PredictProbability(CreditRisk)

FROM CreditRisk **PREDICTION JOIN** NewCustomers

ON CreditRisk.Gender=NewCustomer.Gender

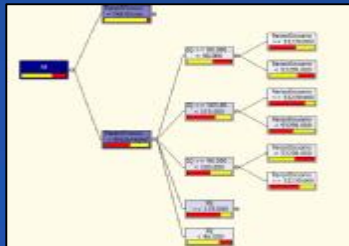
AND CreditRisk.Income=NewCustomer.Income

AND CreditRisk.Profession=NewCustomer.Profession

您的潜力. 我们的动力

Microsoft
微软(中国)有限公司

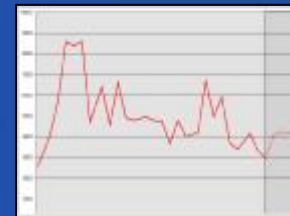
丰富的算法集合



决策树



聚类



时间序列

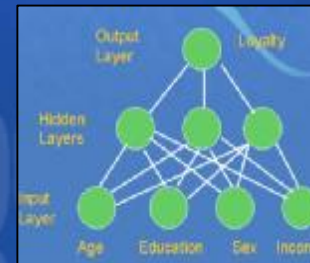
Naïve 贝叶斯



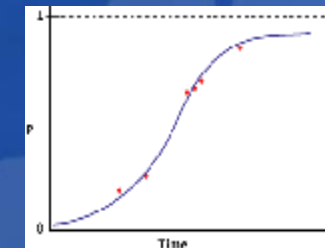
序列聚类



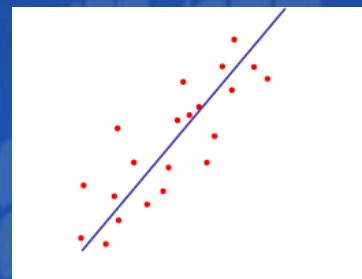
关联



神经网络



逻辑回归



线性回归

When upgrading to Microsoft® SQL Server™ 2005, you can upgrade servers in your organization at a time. However, when servers are used for **Microsoft®**, you must upgrade the **Microsoft®** Publisher, server, and data Subscriber. Upgrading servers one at a time following this is recommended when a large number of Publishers and Subscribers exist because you can upgrade data even though servers are running different versions of SQL Server. You can upgrade Publishers and Subscribers with server running versions of SQL Server 2000, and in Subscribers created in SQL Server 6.5 or SQL Server 7.0.

When using transactional **Microsoft®**, you can upgrade subscribers before the publisher, using immediate updating with **Microsoft®** or transactional **Microsoft®**. There are complete recommendations in this topic under Upgrading and Immediate Updating.

You can upgrade **Microsoft®** servers running SQL Server 6.5 or SQL Server 7.0 to SQL Server 2005. You can upgrade SQL Server 6.5, you do not need to upgrade to SQL Server 7.0 to upgrade to SQL Server 2005.

IMPORTANT: When upgrading servers configured for **Microsoft®** to SQL Server 2005, the compatibility level must be set to 70 (version 7.0 compatibility) or later. If you have a running in SQL Server 6.5 or an earlier compatibility level, temporarily change them during the upgrade process.

When the Publisher or Subscriber is running in 6.5 or an earlier compatibility level and is SQL Server 2000, error 28044 will be raised stating that the operation is successful. SQL Server version 7.0 or SQL Server 2005.

For more information about setting the backward compatibility level, see **SQL Server 2005: Backward Compatibility**.

If you are upgrading **Microsoft®** on a Publisher cluster, you must produce the articles in before upgrading, considering the previous installation means that you must delete all articles **Microsoft®** and reconfigure it after upgrading to SQL Server 2005. This will not require you upgrading SQL Server 2005 to future releases.

文本挖掘

您的潜力. 我们的动力

Microsoft
微软(中国)有限公司

tNaive Bayes 贝叶斯算法

- 概念篇
- 参数篇
- 结果展现
- 适用场景

您的潜力. 我们的动力

Microsoft
微软(中国)有限公司

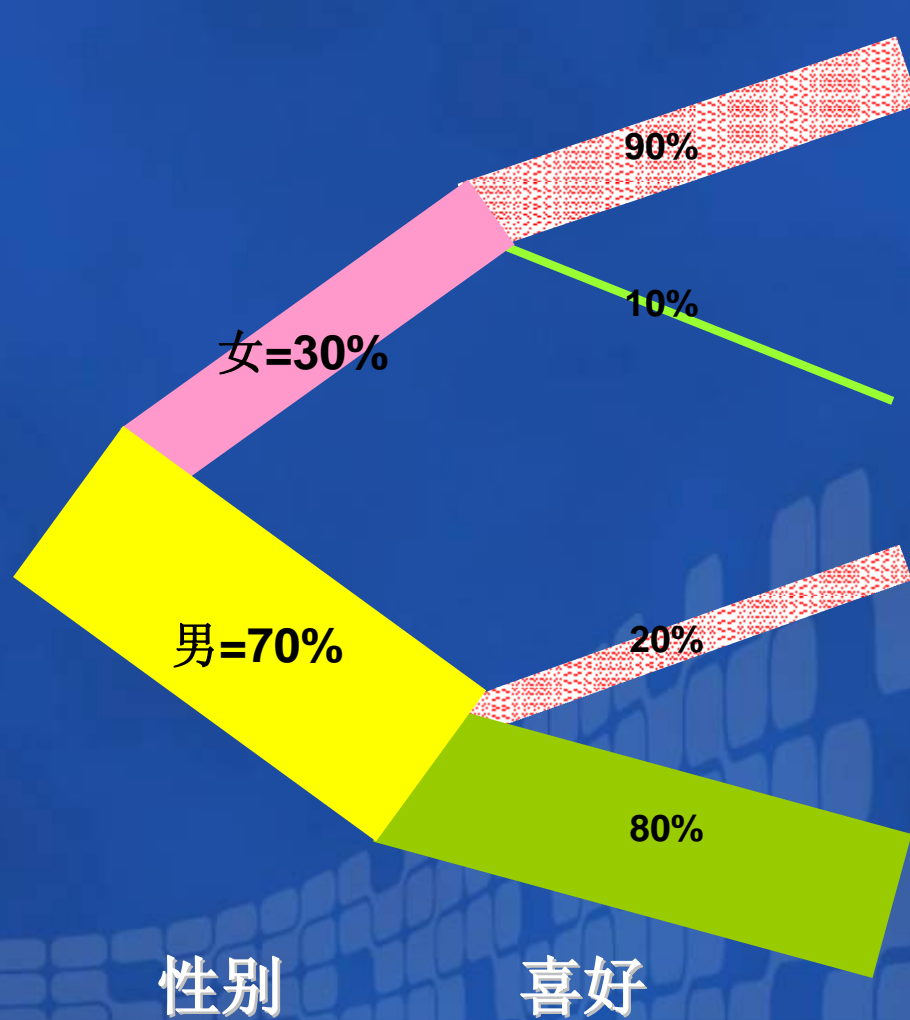
NB概念篇

- 针对分类问题，快速搭建挖掘模型，进行预测
- 假设各种属性之间互相没有影响。
- 可以帮助我们迅速理解数据的特点

您的潜力. 我们的动力

Microsoft
微软(中国)有限公司

超级女声



张靓颖

李宇春

.27

玉米: .41

.03

.14

凉粉: .59

.56

数学基础

- 已知: 70%的男性, 59%的凉粉, 80%的男性是凉粉
- 问题: 百分之多少的凉粉是男性?
- $P(H)$: H的概率
- $P(H|E)$: 在E的情况下H的概率

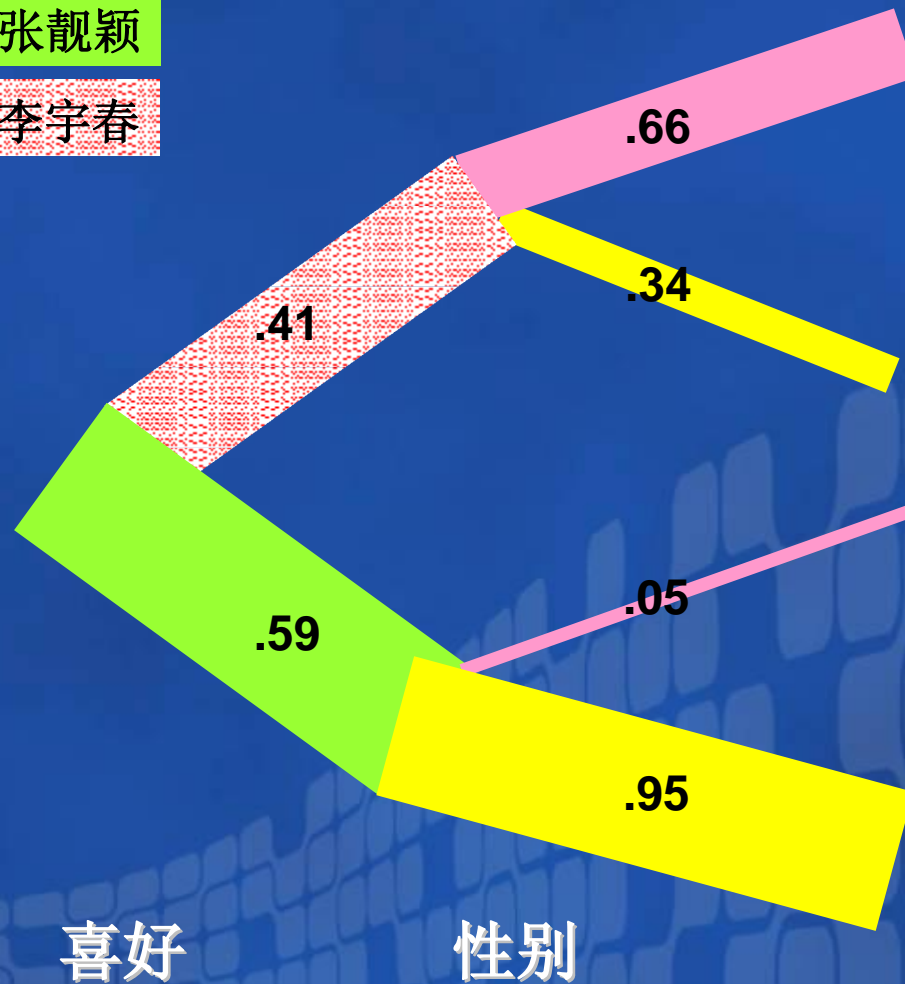
$$P(H | E) = \frac{P(E | H) * P(H)}{P(E)}$$

超级女声：预测投票者性别

张靓颖

李宇春

预测投票者的性别，准确度大大加强



您的潜力. 我们的动力

Microsoft
微软(中国)有限公司

NB Demo 1

- 超级女声
 - 建模
 - 处理
 - 预测

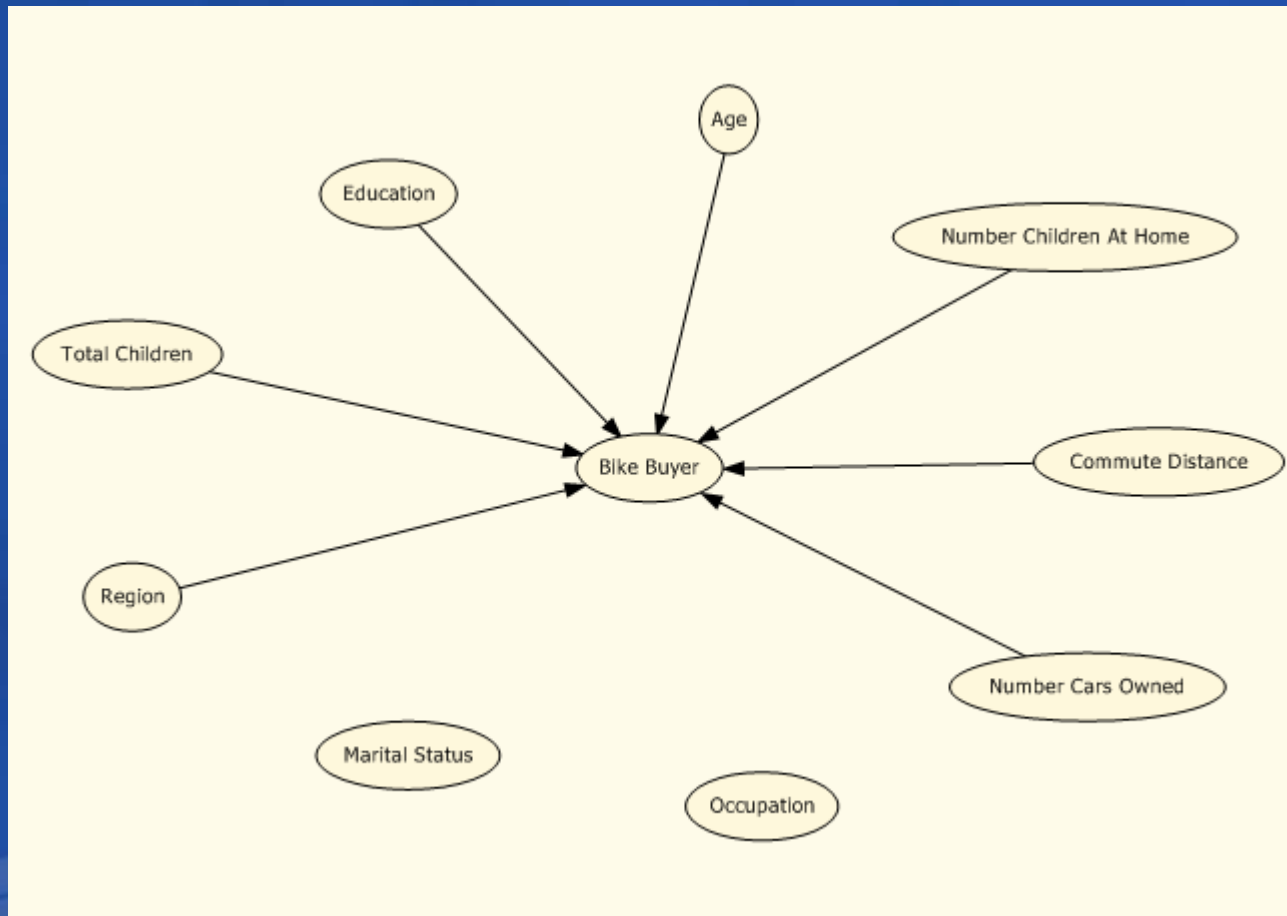
NB Demo 2

- 已知
 - 性别
 - 年龄
 - 交通距离
 - 收入
 - 汽车数目
 - 子女数目
- 预测
 - 潜在的自行车客户

您的潜力. 我们的动力

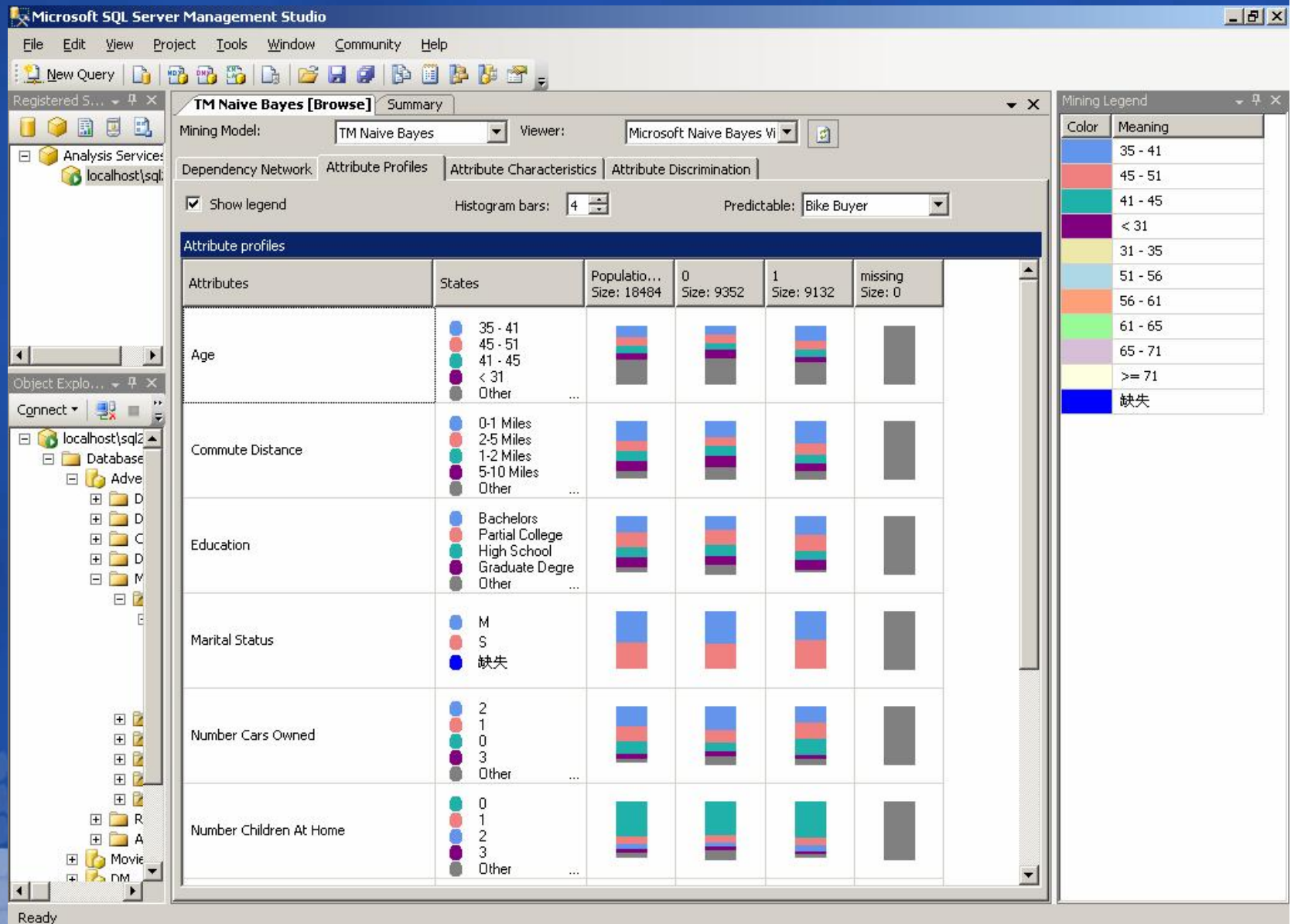
Microsoft
微软(中国)有限公司

Dependency Network



Microsoft®
微软(中国)有限公司

Attribute Profiles



Attribute Characteristics

您的潜力. 我们的动力

Microsoft
微软(中国)有限公司

Microsoft SQL Server Management Studio

File Edit View Project Tools Window Community Help

Registered S... TM Naive Bayes [Browse] Summary

Mining Model: TM Naive Bayes Viewer: Microsoft Naive Bayes Vi

Dependency Network Attribute Profiles Attribute Characteristics Attribute Discrimination

Attribute: Bike Buyer Value: 0

Characteristics for 0

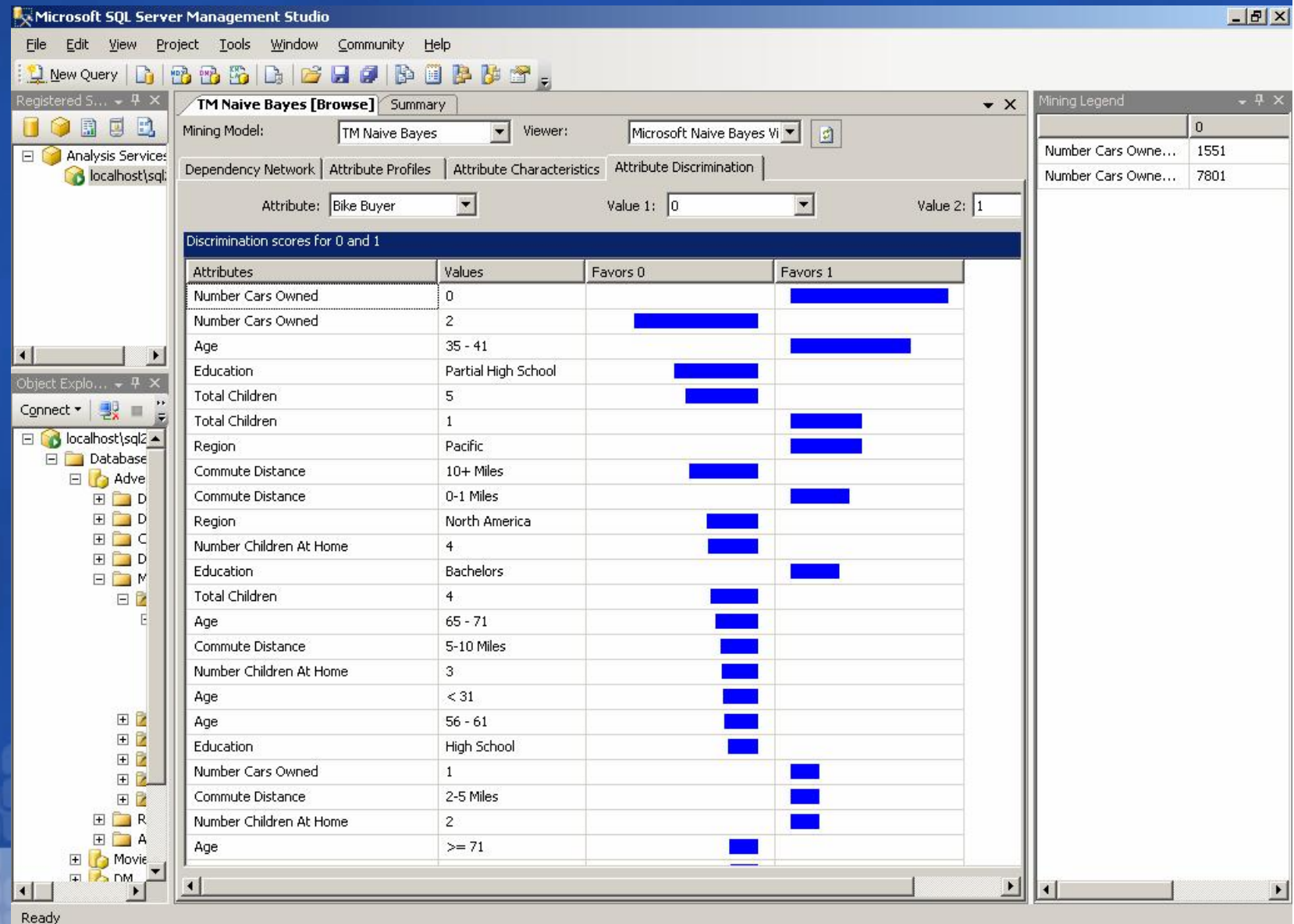
Attributes	Values	Probability
Number Children At Home	0	
Marital Status	M	
Region	North America	
Marital Status	S	
Number Cars Owned	2	
Commute Distance	0-1 Miles	
Region	Europe	
Occupation	Professional	
Education	Partial College	
Total Children	0	
Education	Bachelors	
Occupation	Skilled Manual	
Number Cars Owned	1	
Education	High School	
Commute Distance	5-10 Miles	
Total Children	2	
Commute Distance	1-2 Miles	
Occupation	Management	
Commute Distance	10+ Miles	
Number Cars Owned	0	
Education	Graduate Degree	
Age	35 - 41	
Age	45 - 51	
Total Children	1	

msdn

Ready

Microsoft®
微软(中国)有限公司

msdn



NB 参数

- Maximum_Input_Attributes
- Maximum_Output_Attributes
- Maximum_States
- Minimum_Dependency_Probability

适用场景

- 分类
 - 把新的案例分配给已定义类别
- 典型问题:
 - 银行贷款申请分类
 - 客户分类
 - 迅速对数据获得基本的理解

您的潜力. 我们的动力

Microsoft
微软(中国)有限公司

t 决策树 Decision Trees

- 概念篇
 - 线性回归
 - 回归树
- 参数篇
- 结果展现
- 适用场景

决策树概念篇

- 可以预测离散的, 或者连续的数值
- 把已知条件 (不论是离散还是连续) 自动分解为多个离散的类别。
- 挖掘的结论易于理解
- 初始状态是一个大的空间, 挖掘的过程是递归分区 – 不断分割

案例

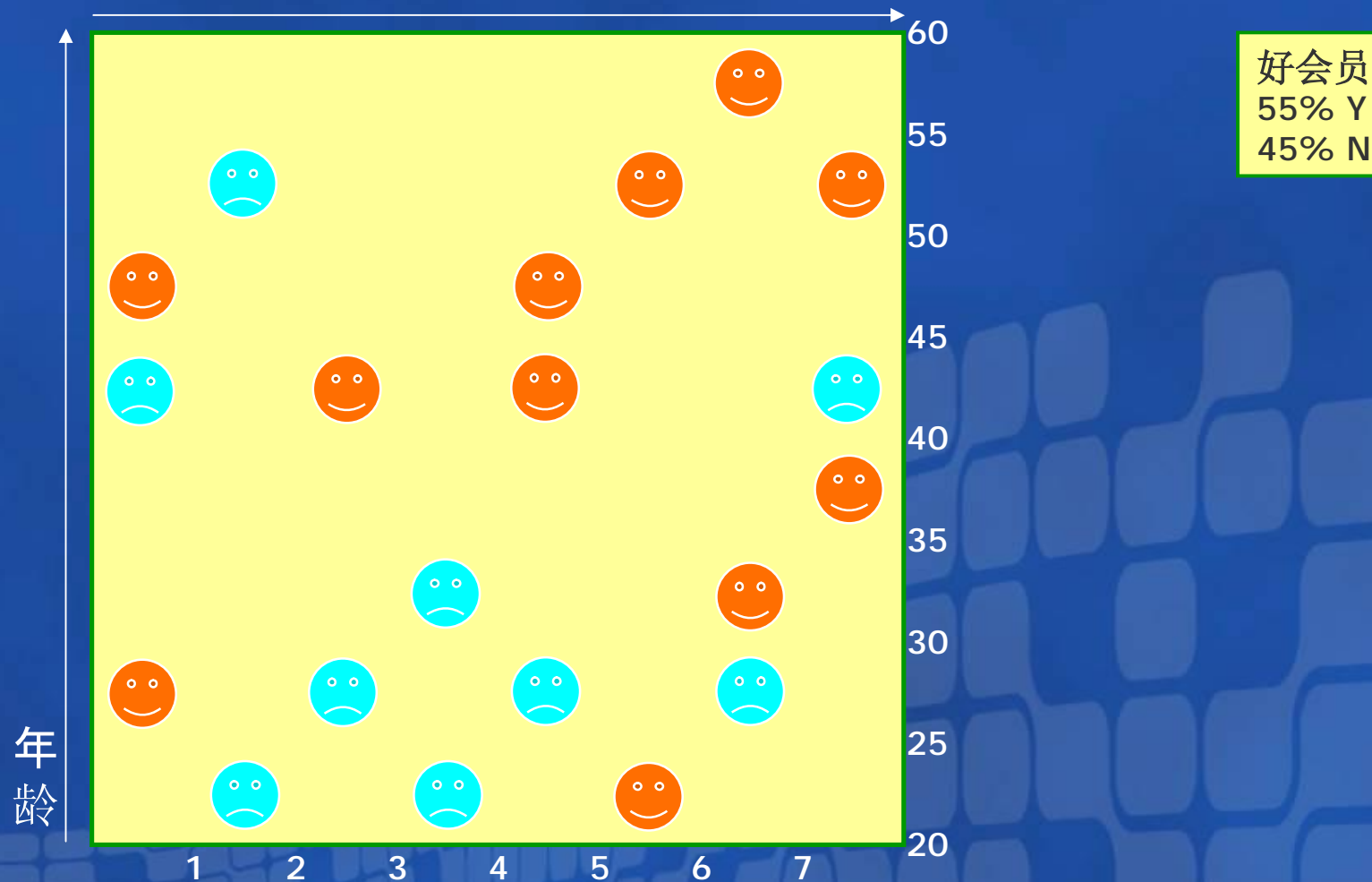
- 我们有大量的会员
 - 年龄在 20 – 60 岁
 - 月薪在 0 – 8000 元
- 55% 的被我们认可为忠实会员（好会员）
- 里面潜在的规律是什么？

您的潜力. 我们的动力

Microsoft
微软(中国)有限公司

谁是我们的忠实会员？

月薪（千元）



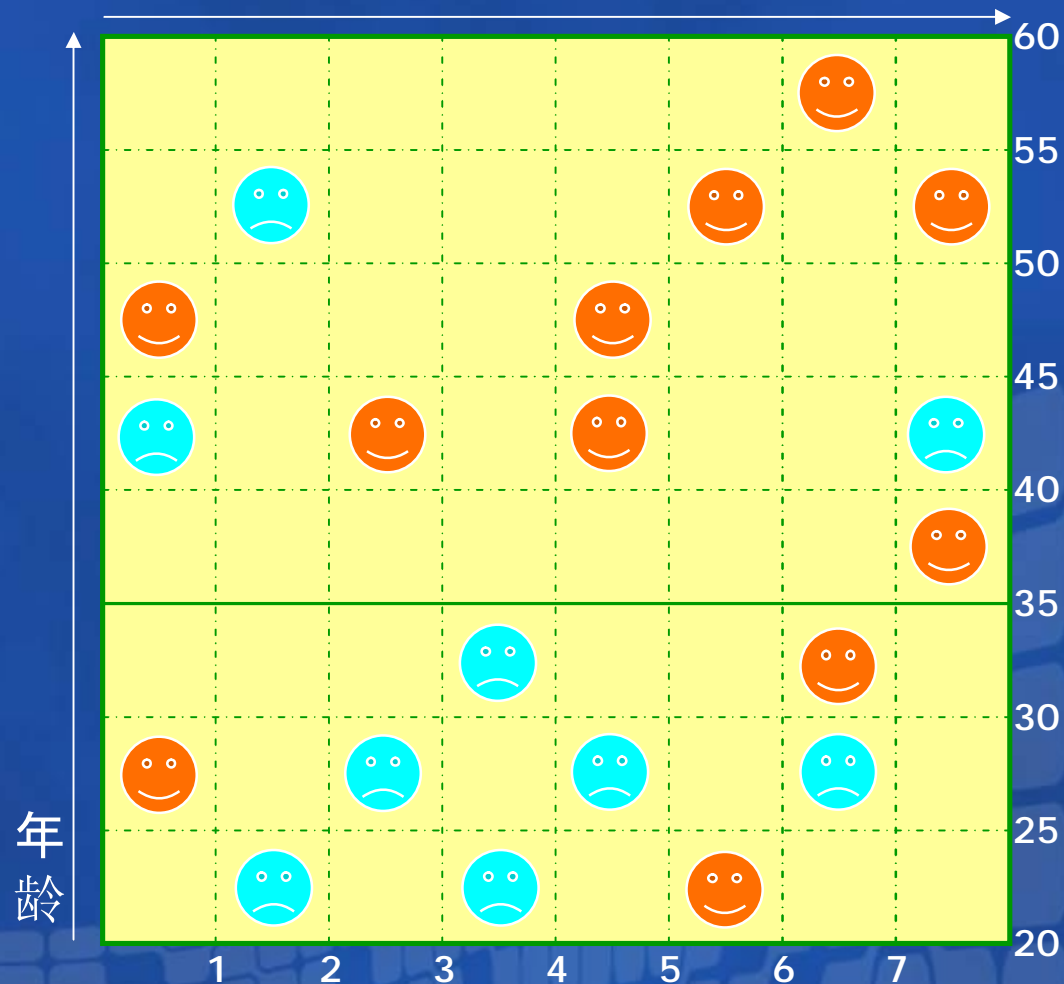
好会员
55% Y
45% N

您的潜力. 我们的动力

Microsoft
微软(中国)有限公司

谁是我们的忠实会员？

月薪（千元）



好会员
55% Y
45% N

35+

年龄

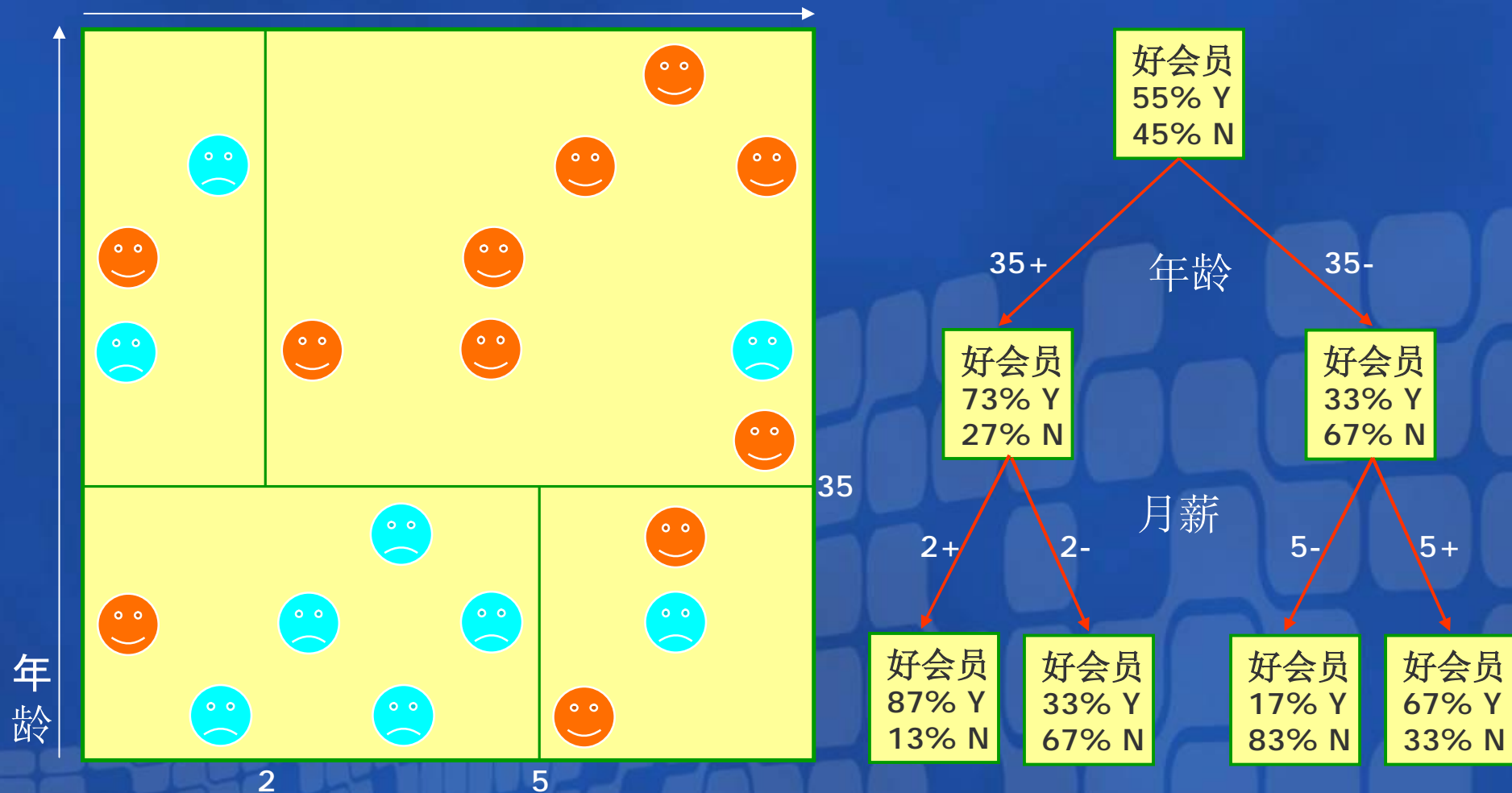
35-

好会员
73% Y
27% N

好会员
33% Y
67% N

谁是我们的忠实会员？

月薪（千元）



线性回归

- 线性回归算法把变量表示为线性函数, 例如
 - 销售额 = $a + b * \text{时间}$
- 多元线性回归可以描述多个维度, 例如
 - $Y = a + b_1 * X_1 + b_2 * X_2 + \dots$

决策树参数

- Maximum_Input_Attributes
- Maximum_Output_Attributes
- Score_Method
- Split_Method
- Minimum_Support
- Complexity_Penalty
- Forced_Regressor

决策树 Demo 2

- 已知
 - 性别
 - 年龄
 - 交通距离
 - 收入
 - 汽车数目
 - 子女数目
 - 预测
 - 潜在的自行车客户
- 已知
 - 性别
 - 年龄
 - 交通距离
 - 购买自行车情况
 - 汽车数目
 - 子女数目
 - 预测
 - 收入

决策树结果展现

- 被预测的是离散属性:
 - 分支条件
 - 预测数值用彩色水平条描述, 根据概率大小排序
- 被预测的是连续属性:
 - 菱形图
 - 回归等式
- 关联网络
 - 展现挖掘模型中各种属性的预测能力之间的关联

Salary Predict [Online] Targeted Mailing [Online] Start Page

Mining Structure Mining Models Mining Model Viewer Mining Accuracy Chart Mining Model Prediction

Mining Model: TM Decision Tree Viewer: Microsoft Tree Viewer

Decision Tree Dependency Network

Tree: Bike Buyer Default Expansion: 3 Levels

Histograms: 6 Background: All Cases Show Level 1 Level 11

全部

Number Cars Owned = 2

Yearly Income < 26000

Yearly Income >= 58000 and < 106000

Yearly Income >= 26000 and < 58000

Number Cars Owned = 4

Commute Distance 不等于 0-1 Mil...

Commute Distance = 0-1 Miles

Solution Explorer

- Adventure Works DW(DYANGNOTI)
- Data Sources
- Adventure Works DW
- Data Source Views
- Cubes
- Dimensions
- Mining Structures
 - Targeted Mailing
 - Market Basket
 - Sequence Clustering
 - Forecasting
 - Customer Mining
 - Salary Predict
- Roles

Solution Explorer Class View

Mining Legend

High Low

Total Cases: 6457

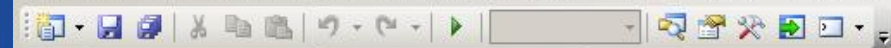
Value	Cases	Probabil
<input checked="" type="checkbox"/> 0	3868	59.89%
<input checked="" type="checkbox"/> 1	2589	40.09%
<input checked="" type="checkbox"/> 缺失	0	0.01%

Number Cars Owned = 2

Properties Mining Legend

Error List Output Find Symbol Results

Ready



Mining Model: y Target Mail

Viewer: Microsoft Tree Viewer

Decision Tree Dependency Network



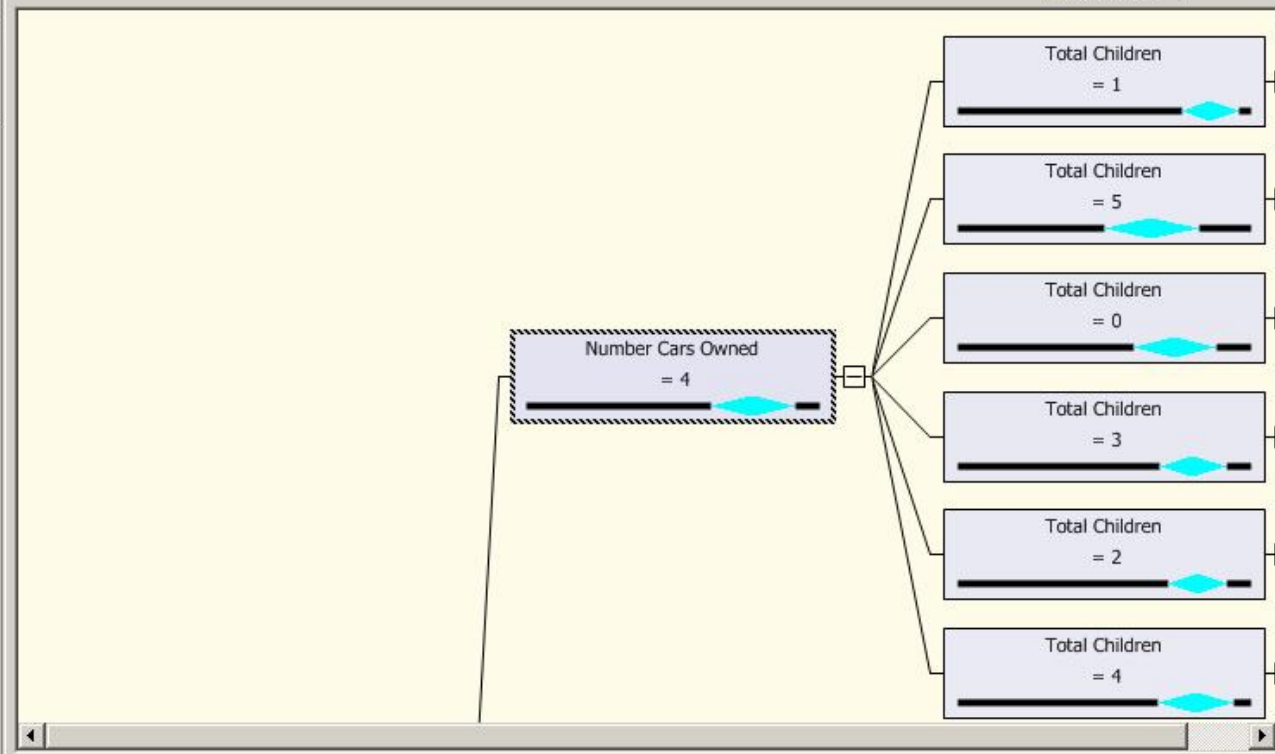
Tree: Yearly Income

Default Expansion: 3 Levels

Histograms: 6

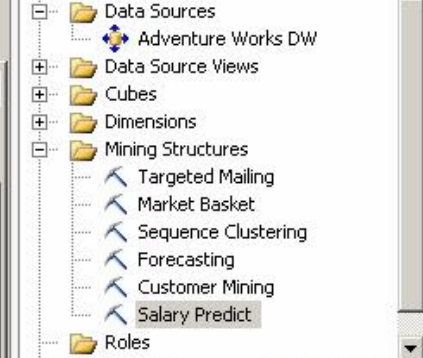
Background: All Cases

Show Level 1 Level 10



Error List Output Find Symbol Results

Ready



Term		Coeffici.
Age	*	76591...
		745.392

Number Cars Owned = 4
Existing Cases: 1261
Missing Cases: 0
Yearly Income = 110,428.290+745.392*
(Age-45.395)



Salary Predict [Online] Targeted Mailing [Online] Start Page

Mining Structure Mining Models Mining Model Viewer Mining Accuracy Chart Mining Model Prediction

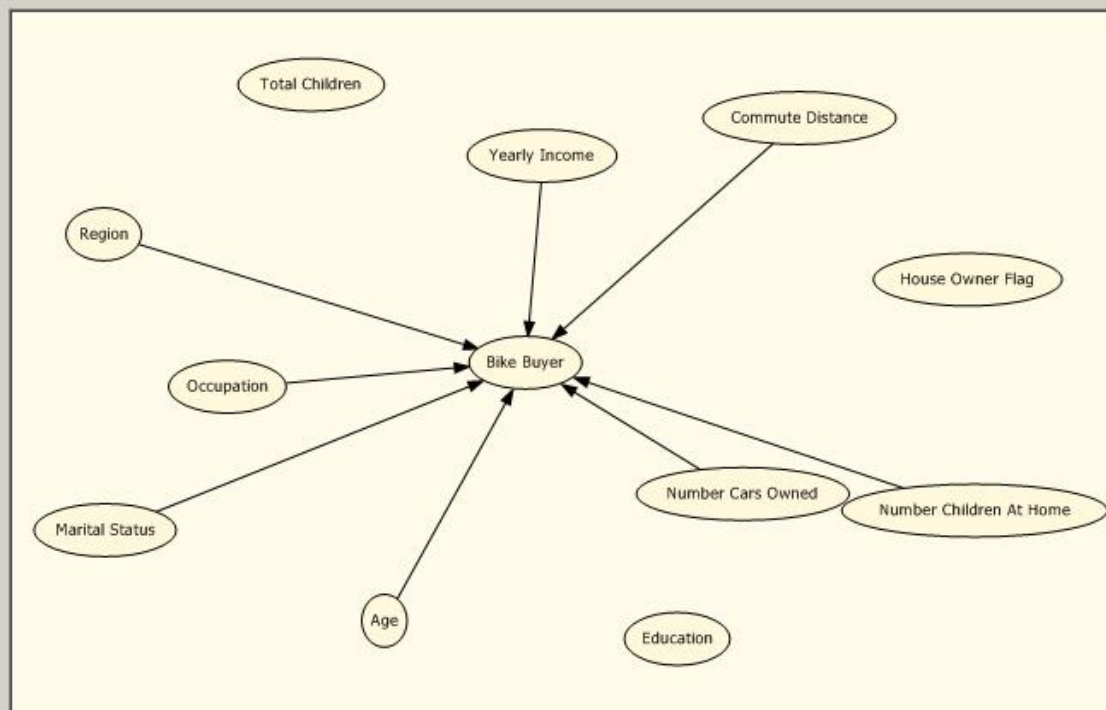
Mining Model: TM Decision Tree

Viewer: Microsoft Tree Viewer

Decision Tree Dependency Network

☐ Show long name

All Links



Select a node in the network to highlight its dependencies.

Strongest Links

Selected node

Node it predicts

Node that predicts it

Predicts both ways

Error List Output Find Symbol Results

Ready

Solution Explorer



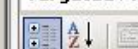
Adventure Works DW(DYANGNOTI

- Data Sources
 - Adventure Works DW
- Data Source Views
- Cubes
- Dimensions
- Mining Structures
 - Targeted Mailing
 - Market Basket
 - Sequence Clustering
 - Forecasting
 - Customer Mining
 - Salary Predict
- Roles

Solution Explorer Class View

Properties

Targeted Mailing MiningStructure



Description

ID Targeted Mailing

Name Targeted Mailing

Misc

CacheMode KeepTrainingCases

[Add a Column](#), [Add a Nested Table](#), [Create a Related Mining Model](#), [Process All](#)

Name

Specifies the name of the object.

决策树的缺点

- 过多的细节末梢
 - 可以事先限制
 - 可以事后切除
- 数据量过大时, 性能会有问题

您的潜力. 我们的动力

Microsoft
微软(中国)有限公司

适用场景

- 分类和预测
- 典型问题
 - 预测潜在客户
 - 评估客户风险
 - 找到决策规则

您的潜力. 我们的动力

Microsoft
微软(中国)有限公司

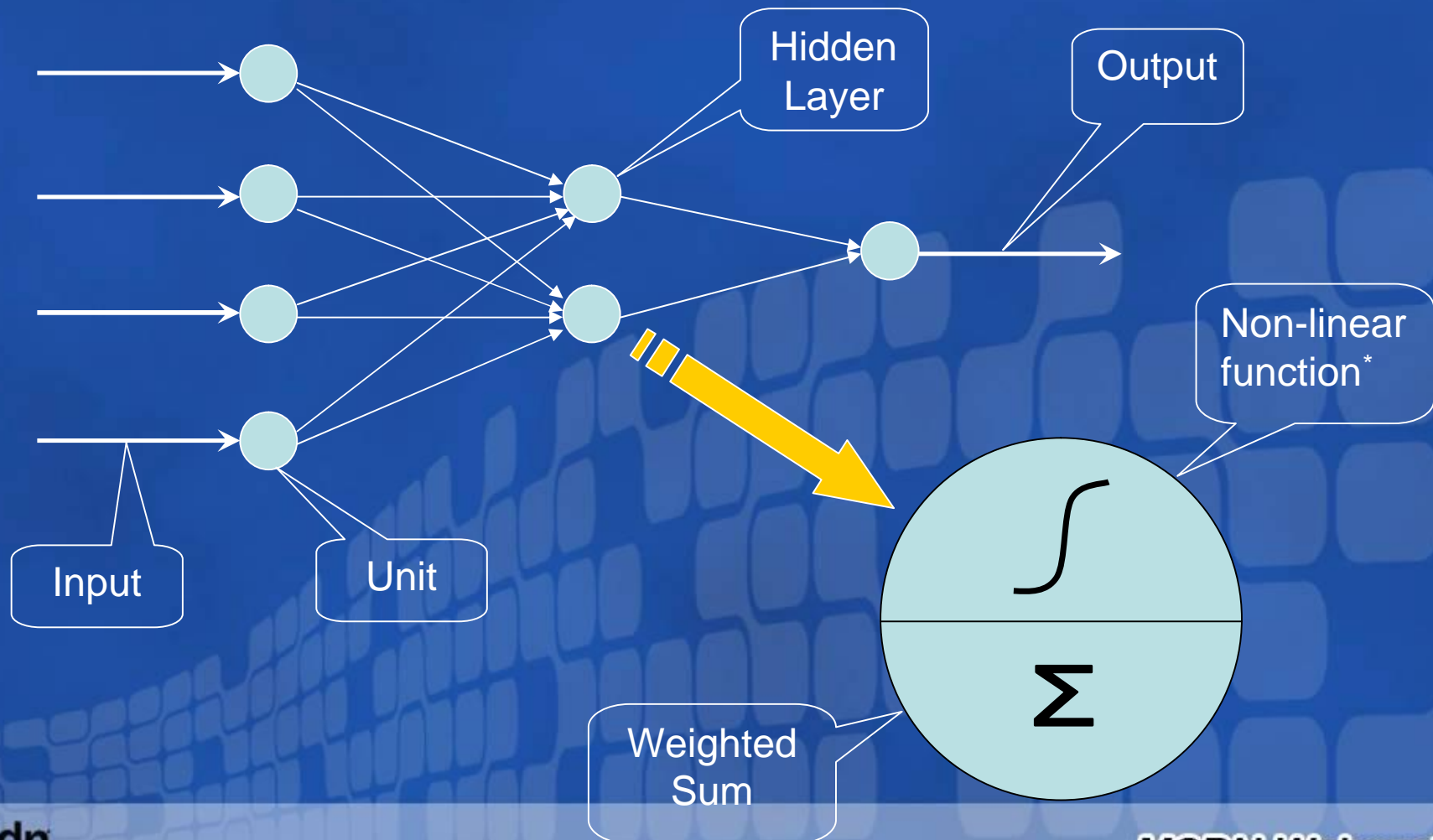
t 神经网络

- 概念篇
 - 逻辑回归
- 参数篇
- 结果展现
- 适用场景

NN 概念

- 模拟人类大脑
 - 通过学习获得知识
 - 知识存储在神经元网络里
- 可以找出全部可能的数据关系
 - 但是它很慢!
- 有可能处理复杂情况的一种算法

神经网络训练过程



Backpropagation

- Training a neural network is setting the best weights on the inputs of each of the units
- The backpropagation process:
 - Get a training example and calculate outputs
 - Calculate the error – the difference between the calculated and the expected (known) result
 - Adjust the weights to minimize the error

逻辑回归（非线性回归）

- 最简单的神经网络（输入神经元直接连接输出神经元，没有中间的隐藏层）就是一个逻辑回归方程

神经网络参数

- Maximum_Input_Attributes
- Maximum_Output_Attributes
- Hidden_Node_Ratio
- Holdout_Percentage
- Holdout_Seed
- Maximum_States
- Sample_Size

NN结果展现

- 并非像决策树那样直观
- 显示了每一种输入值和被预测值之间的影响。
- 可以看不同被预测值的差异

适用场景

- 类似决策树
 - 分类
 - 预测
- 理解起来更复杂, 因此, 不如决策树那样常见

您的潜力. 我们的动力

Microsoft
微软(中国)有限公司

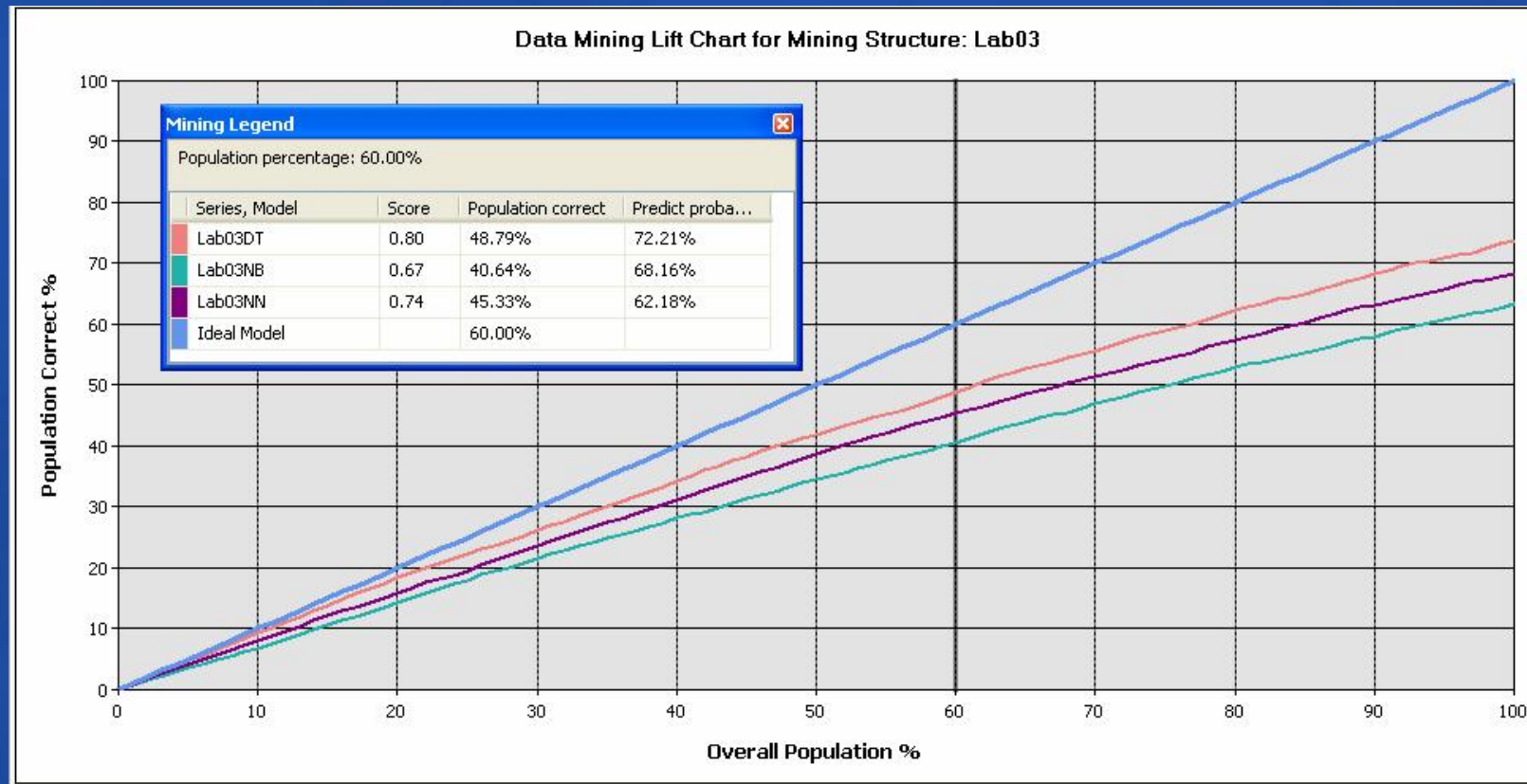
比较挖掘的准确度

- Lift Chart
- Profit Chart
- Classification Matrix

Lift Chart

您的潜力. 我们的动力

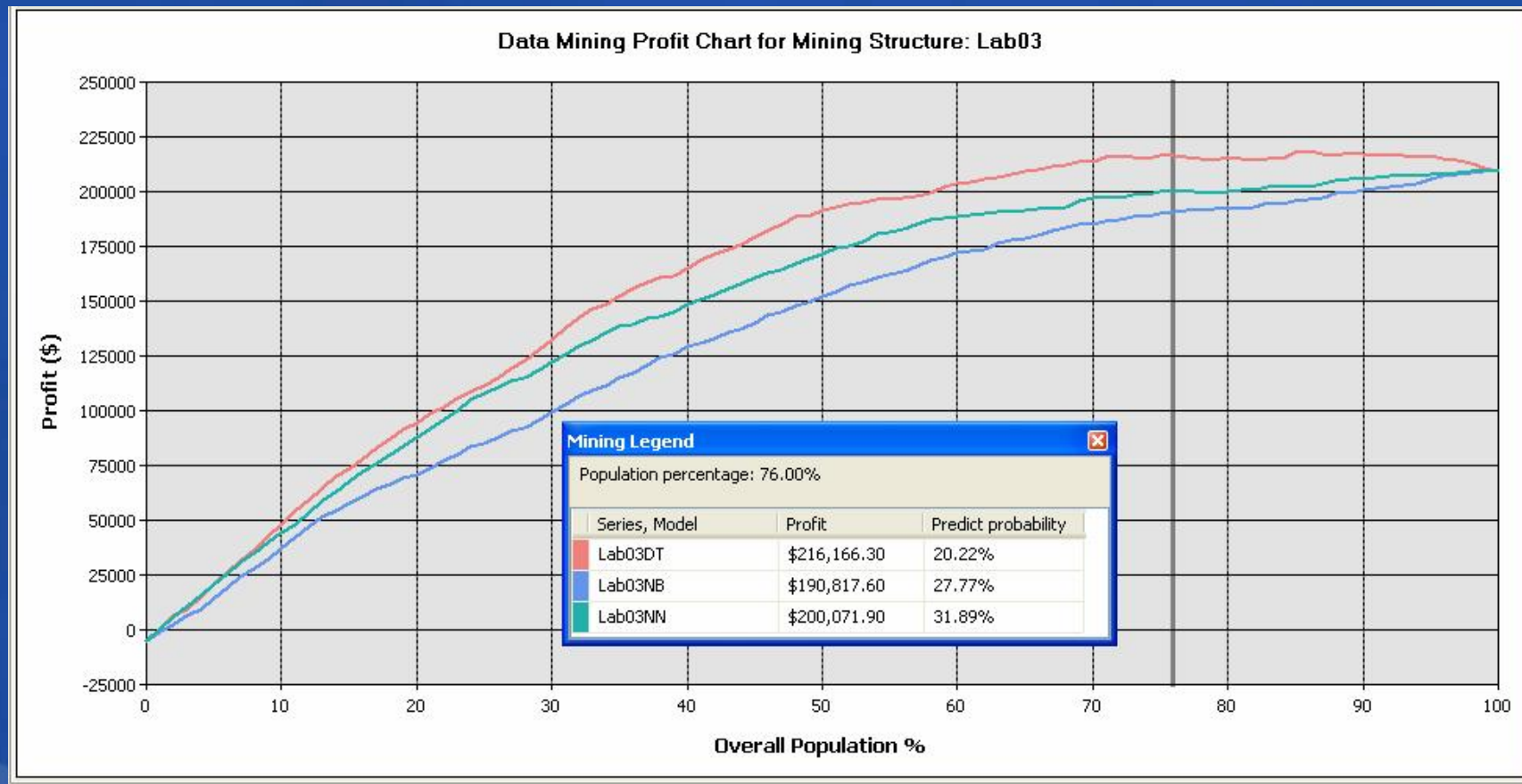
Microsoft
微软(中国)有限公司



您的潜力. 我们的动力

Microsoft
微软(中国)有限公司

Profit Chart



Classification Matrix



Columns of the classification matrices correspond to actual values; rows correspond to predicted values

Counts for Lab03DT on [Bike Buyer]:

Predicted	0 (Actual)	1 (Actual)
0	1475	539
1	441	1273

Counts for Lab03NB on [Bike Buyer]:

Predicted	0 (Actual)	1 (Actual)
0	1236	684
1	680	1128

Counts for Lab03NN on [Bike Buyer]:

Predicted	0 (Actual)	1 (Actual)
0	1261	528
1	655	1284

t Review


- SQL2005数据挖掘概述
- 贝叶斯 (Naive Bayes)
- 决策树 (Decision Trees)
 - 线性回归 (Linear Regression)
- 神经网络 (Neural Networks)
 - 逻辑回归 (Logistic Regression)
- 比较挖掘的准确度

获取更多MSDN资源

- **MSDN中文网站**
<http://www.microsoft.com/china/msdn>
- **MSDN中文网络广播**
<http://www.msdnwebcast.com.cn>
- **MSDN Flash**
<http://www.microsoft.com/china/newsletter/case/msdn.aspx>
- **MSDN开发中心**
<http://www.microsoft.com/china/msdn/DeveloperCenter/default.msp>



Question & Answer

如需提出问题，请单击“提问”按钮并在随后显示的浮动面板中输入问题内容。一旦完成问题输入后，请单击“提问”按钮。

 **问题和解答 (无问题)** ▲ ×

在此会议中尚未解答任何问题。

要向演示者提问，请在此处键入问

提问(A)

删除(D)

问题管理器(Q)

您的潜力，我们的动力

Microsoft®
微软(中国)有限公司

Microsoft®

msdn


MSDN Webcasts