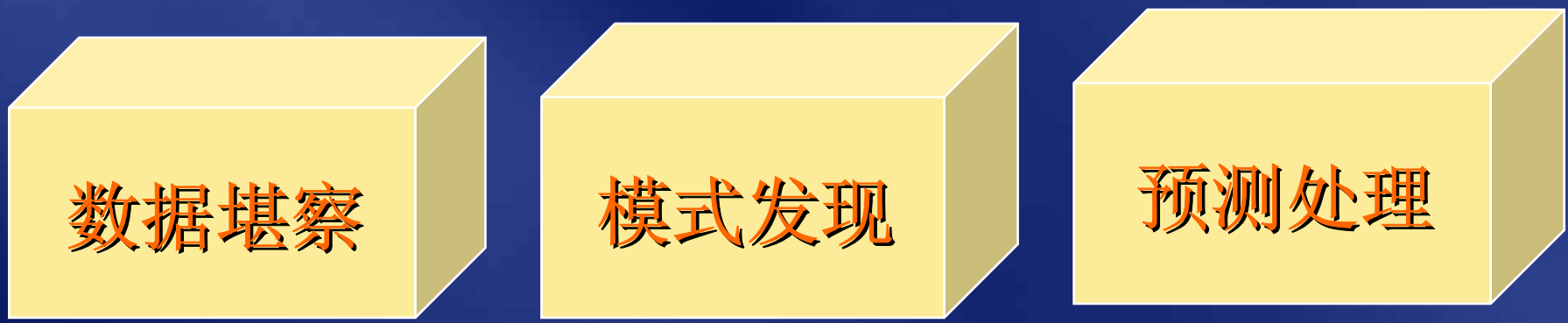
SQL Server 2005 之数据挖掘

唐朝晖
高级项目经理
SQL Server商务智能
Microsoft公司

提纲

- 数据挖掘概述
  - 什么是数据挖掘?
  - 典型商务问题
- SQL Server 2005 系统中的数据挖掘
  - Key messages
  - DMX
  - 新特征介绍
    - 有关工具
    - 分析引擎
    - BI 集成
- 演示
- Q&A

什么是数据挖掘?



典型的商业问题

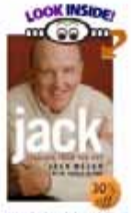
- 客户的信用风险是什么?
- 我的客户是哪些?
- 客户比较倾向于同时买哪些商品?
- 下个月我能销售多少Pepsi可乐?
- 我的商店的潜在客户会有多少?
- 在互联网上“我”的主要客户群又是哪些?
- ...

Amazon.com: Books: Jack: Straight from the Gut - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media

Address http://www.amazon.com/exec/obidos/tg/detail/-/0446528382/qid=1042753337/sr=8-1/ref=sr_8_1/104-2680659-3995951?v=glance&s=books&n=50784 Go Links



Look inside this book

List Price: $29.95

You
Ava

Nee
vary
your

Used & new from $7.

Edition: Hardcover |

See more product

Great Buy

Buy this book with G

Buy To

Buy bo

Customers who

- The Jack Welch
  Legendary Lead
- Execution: The
- First, Break All
  Curt Coffman (
- Jack Welch & T
  Robert Slater (
- Built to Last: S

BOOK INFORMATION
buying info
editorial reviews
customer reviews
look inside

RATE THIS ITEM
I dislike it I love it!
1 2 3 4 5
I own it
Submit
Edit your ratings

Favorite Magazines!

Subscribe to other Business & Finance magazines.

Visit the DVD

Sign in to turn on 1-Click ordering.

Automotive.com | Get a Quote. - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media

Address http://www.automotive.com/redir/link.html?src_id=1316&kw_sqid=101860&kw_am=m Go Links

AUTOMOTIVE.com

HOME | HOW IT WORKS | TESTIMONIALS | GET A QUOTE

" I got the car I wanted at a price I could **afford**. "

Jenna T.- TX

Testimonials »»

DRIVEN BY AUTOMOTIVE.com

- How it Works
- Testimonials
- Get a Quote

Get a Quote.

Get a price in just **minutes** !

Make Ford

Model Explorer

Service Available:

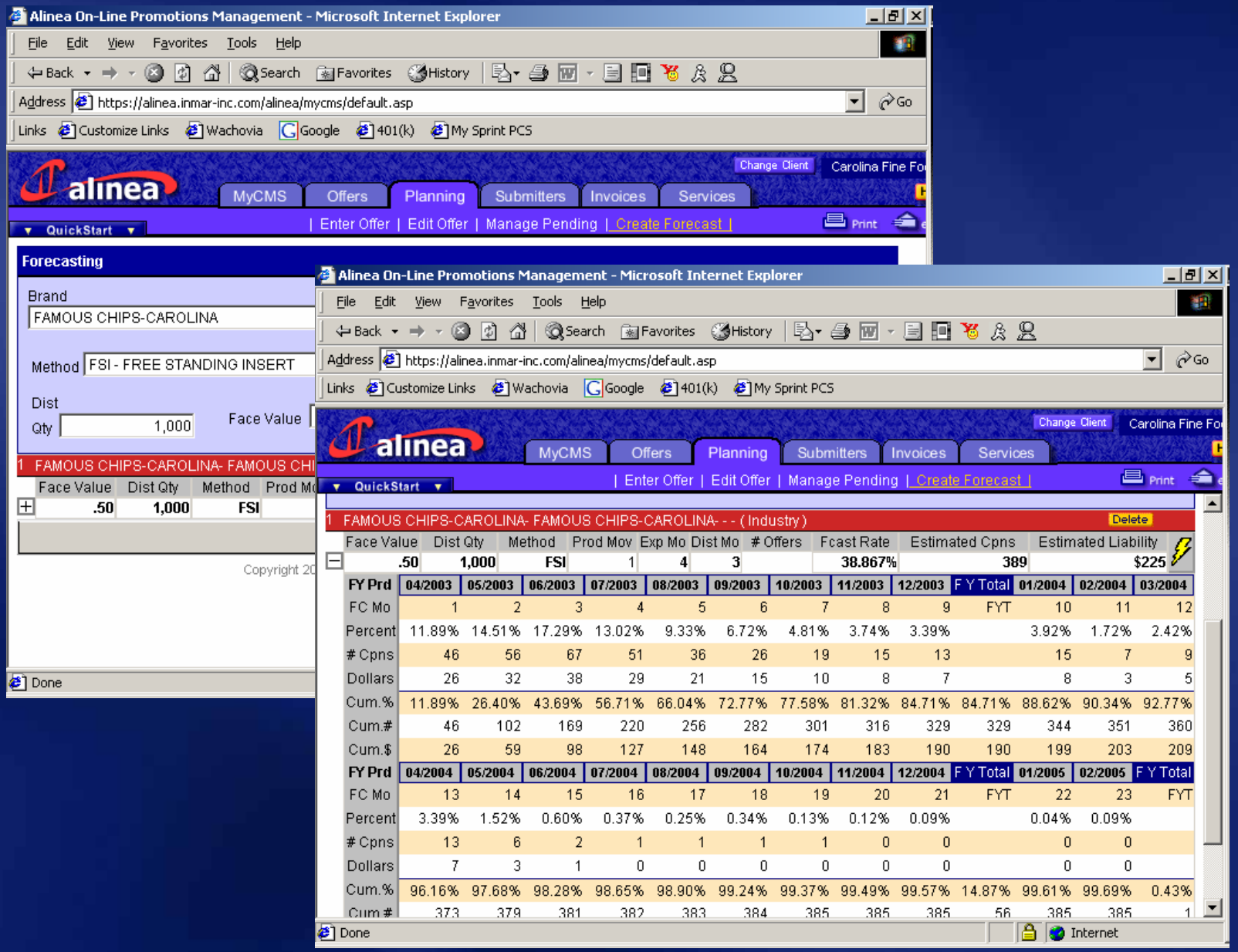
Zip* 98007

Get a Price

Our FREE **New Car Quote** allows you to select the exact car you want, get invoice pricing and receive the best possible price from an accredited Dealer.

Questions ? Please contact us at info@automotive.com. All rights reserved © Automotive.com

Internet


Alinea On-Line Promotions Management - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address https://alinea.inmar-inc.com/alinea/mycms/default.asp
Go
Links Customize Links Wachovia Google 401(k) My Sprint PCS
alinea
Change Client
Carolina Fine Fo
MyCMS Offers Planning Submitters Invoices Services
QuickStart
Enter Offer | Edit Offer | Manage Pending | Create Forecast |
Print
Forecasting
Brand
FAMOUS CHIPS-CAROLINA
Method
FSI - FREE STANDING INSERT
Dist Qty
1,000
Face Value
1 FAMOUS CHIPS-CAROLINA- FAMOUS CHI
Face Value Dist Qty Method Prod Mo
.50 1,000 FSI
Copyright 20
Done
Alinea On-Line Promotions Management - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address https://alinea.inmar-inc.com/alinea/mycms/default.asp
Go
Links Customize Links Wachovia Google 401(k) My Sprint PCS
alinea
Change Client
Carolina Fine Fo
MyCMS Offers Planning Submitters Invoices Services
QuickStart
Enter Offer | Edit Offer | Manage Pending | Create Forecast |
Print
1 FAMOUS CHIPS-CAROLINA- FAMOUS CHIPS-CAROLINA- - - (Industry)
Delete
Face Value Dist Qty Method Prod Mov Exp Mo Dist Mo # Offers Fcast Rate Estimated Cpns Estimated Liability
.50 1,000 FSI 1 4 3 38.867% 389 $225
FY Prd 04/2003 05/2003 06/2003 07/2003 08/2003 09/2003 10/2003 11/2003 12/2003 F Y Total 01/2004 02/2004 03/2004
FC Mo 1 2 3 4 5 6 7 8 9 FYT 10 11 12
Percent 11.89% 14.51% 17.29% 13.02% 9.33% 6.72% 4.81% 3.74% 3.39% 3.92% 1.72% 2.42%
Cpns 46 56 67 51 36 26 19 15 13 15 7 9
Dollars 26 32 38 29 21 15 10 8 7 8 3 5
Cum.% 11.89% 26.40% 43.69% 56.71% 66.04% 72.77% 77.58% 81.32% 84.71% 84.71% 88.62% 90.34% 92.77%
Cum.# 46 102 169 220 256 282 301 316 329 329 344 351 360
Cum.$ 26 59 98 127 148 164 174 183 190 190 199 203 209
FY Prd 04/2004 05/2004 06/2004 07/2004 08/2004 09/2004 10/2004 11/2004 12/2004 F Y Total 01/2005 02/2005 F Y Total
FC Mo 13 14 15 16 17 18 19 20 21 FYT 22 23 FYT
Percent 3.39% 1.52% 0.60% 0.37% 0.25% 0.34% 0.13% 0.12% 0.09% 0.04% 0.09%
Cpns 13 6 2 1 1 1 1 0 0 0 0
Dollars 7 3 1 0 0 0 0 0 0 0 0
Cum.% 96.16% 97.68% 98.28% 98.65% 98.90% 99.24% 99.37% 99.49% 99.57% 14.87% 99.61% 99.69% 0.43%
Cum.# 373 379 381 382 383 384 385 385 385 56 385 385 1
Done
Internet

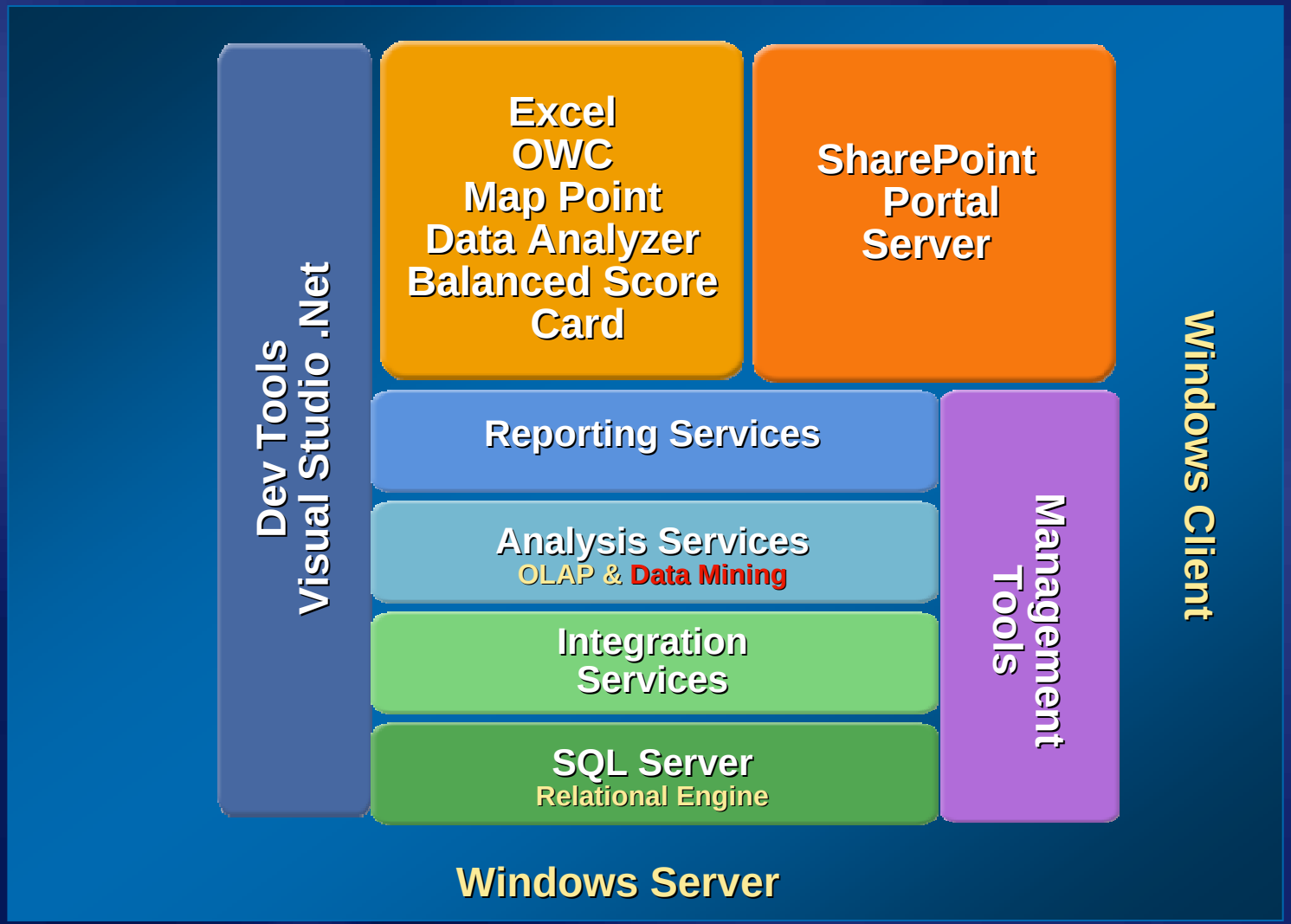
SQL Server 2005 中的数据挖掘

微软的解决方案

进入DM领域
- 创立工业标准
- 整合开发者建议
- 推出产品、步入市场
- V1.0版本中已包括两个著名算法

成为DM领域领导者
- 面向软件开发人员
- 完善行业标准
- 完备的分析功能
- 向企业级应用渗透

数据挖掘 与 Microsoft 商务智能



Key Messages

- 嵌入式数据挖掘
- 集成平台
- 高端分析
- 与ISVs联盟

•从DM 到 LOB 都可嵌入应用

•DM SQL语言- DMX
•XML/A 支持

Key Messages

- 嵌入式数据挖掘
- 集成平台
- 高端分析
- 与ISVs联盟

•与关系数据库(relational)，OLAP，DTS和汇报(reporting)技术的紧密集成

•SQL Server: BI的基础平台

Key Messages

- 嵌入式数据挖掘
- 集成平台
- 高端分析
- 与ISVs联盟

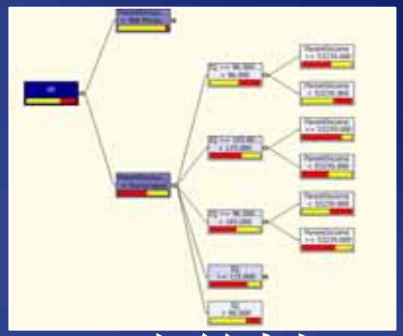
- 主流的数据挖掘算法
- 增强的可视化控件

Key Messages

- 嵌入式数据挖掘
- 集成平台
- 高端分析
- 与ISVs联盟

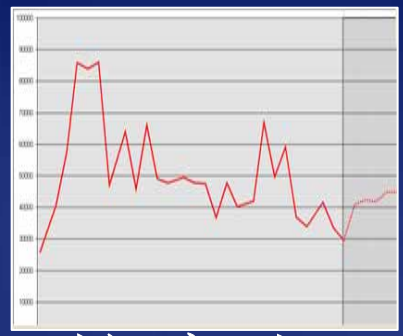
- 倾力关注于市场的拓展
- 基于DMX和XML/A的DM工业标准的最终统一
- 第三方算法Plug-in

完备集算法



决策树

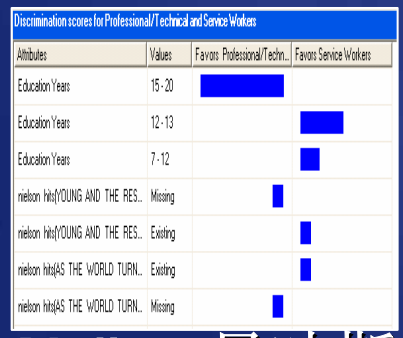
聚类



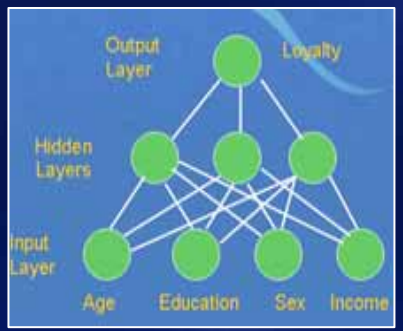
时间序列

序列聚类

关联



Naïve 贝叶斯



神经网络

Text Mining

简单数据

StudentID	Gender	Parent Income	IQ	Encouragement	College Plans
1	Male	23400	120	Not Encouraged	No
2	Female	79200	90	Encouraged	Yes
3	Male	42000	105	Not Encouraged	Yes

复杂数据

Cust ID	Age	Marital Status	IQ	Favorite Movies	
				Title	Score
1	35	M	2	Star Wars	8
				Toy Story	9
				Terminator	7
2	20	S	3	Star Wars	7
				Braveheart	7
				The Matrix	10
3	57	M	2	Sixth Sense	9
				Casablanca	10

API: DMX

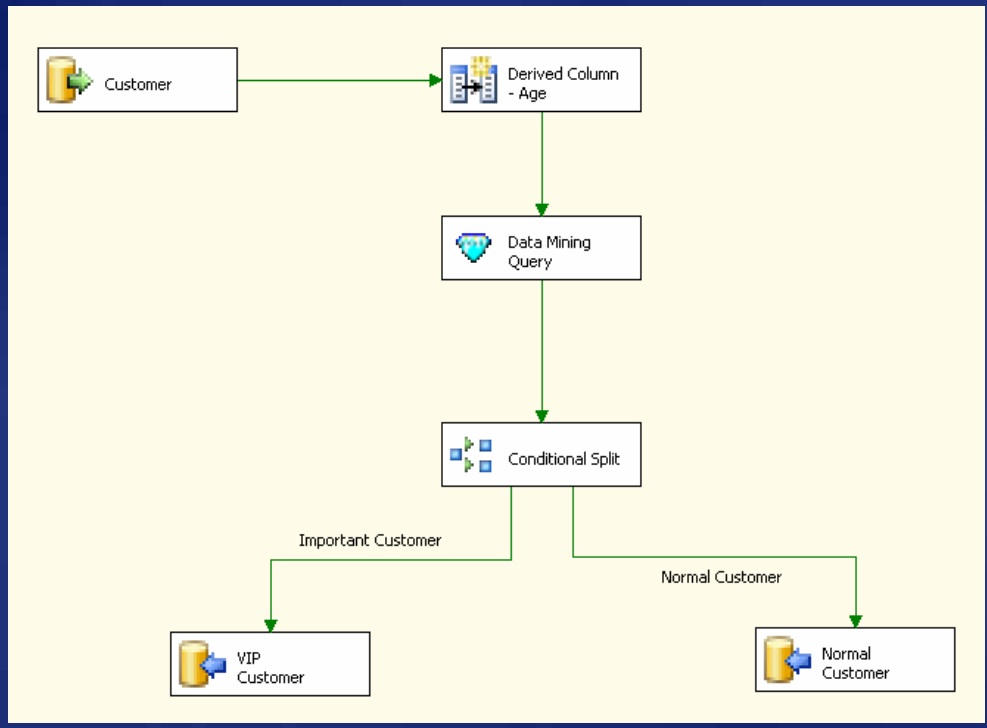
```
CREATE MINING MODEL CreditRisk
(CustID      LONG KEY,
Gender       TEXT DISCRETE,
Income       LONG CONTINUOUS,
Profession   TEXT DISCRETE,
Risk         TEXT DISCRETE PREDICT)
USING Microsoft_Decision_Trees
```

```
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
```

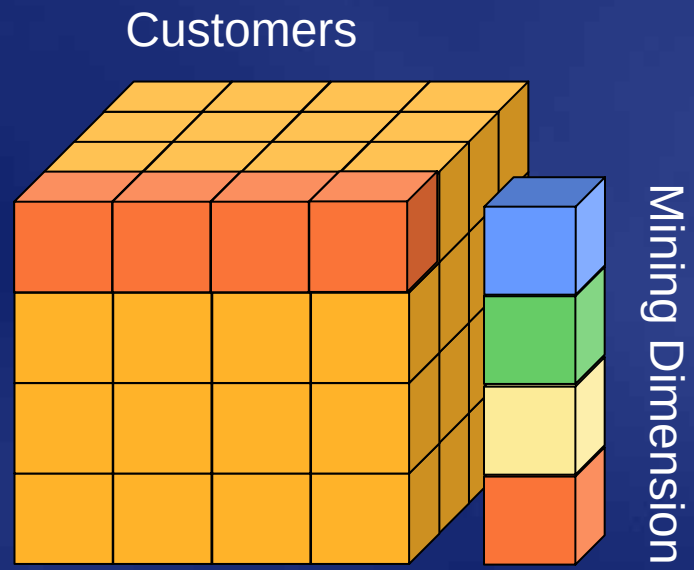
数据挖掘流程

- 与 DTS数据流和任务流的紧密集成
- 数据流
  - 模型训练
  - 预测
  - Text Mining
- 任务流
  - 预测
  - 模型训练



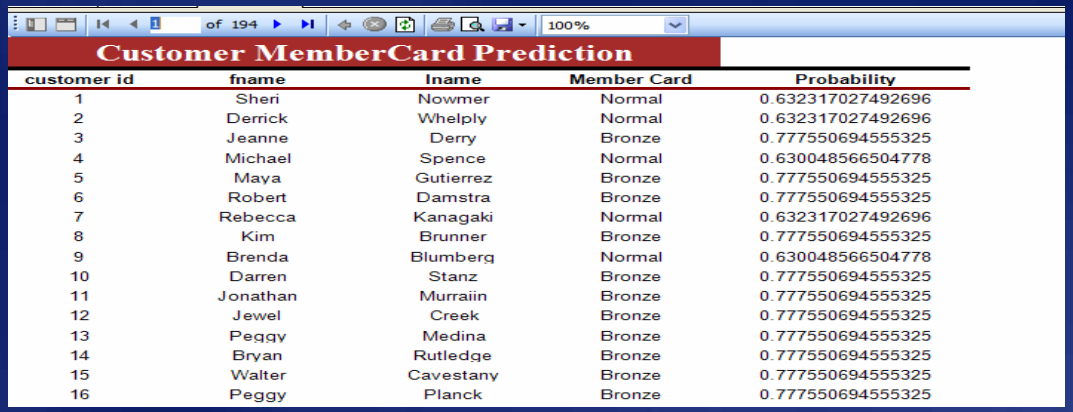
OLAP 挖掘模型增强

- 模型构造上的高灵性
  - 示例：西北地区上月的市场篮子销量分析
  - 预测下一年的销量
- 数据挖掘维度
- 高性能
- 增强的 OLAP 挖掘型向导和编辑器
- 与关系挖掘模型的致性设计



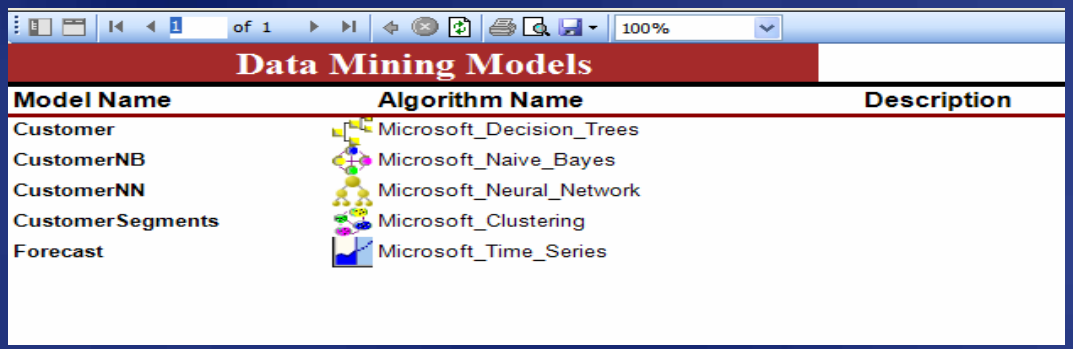
DM报表

- DMX 查询作为 SQL Server 报表数据源
- 报表编译器中嵌入ＤＭＸ查询语言生成器
- Push/Pull mode for DM report delivery



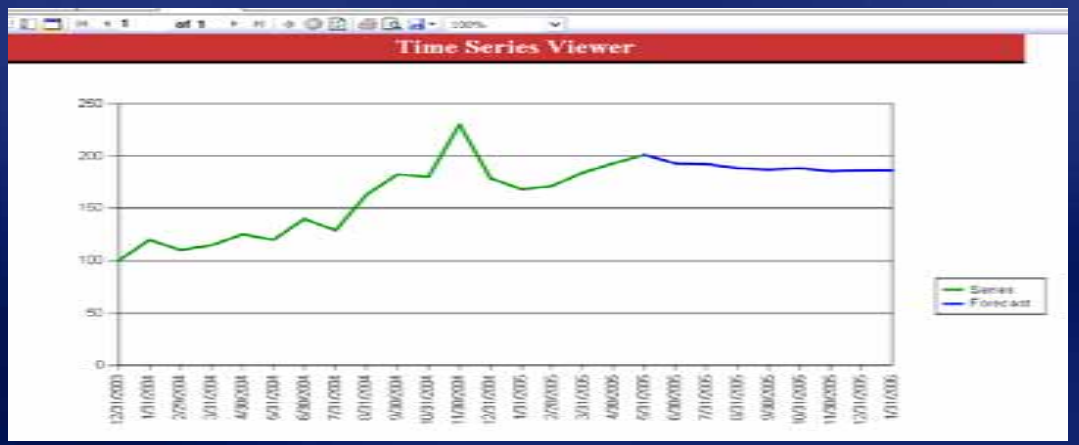
Customer MemberCard Prediction

customer id	fname	lname	Member Card	Probability
1	Sheri	Nowmer	Normal	0.632317027492696
2	Derrick	Whelply	Normal	0.632317027492696
3	Jeanne	Derry	Bronze	0.777550694555325
4	Michael	Spence	Normal	0.630048566504778
5	Maya	Gutierrez	Bronze	0.777550694555325
6	Robert	Damstra	Bronze	0.777550694555325
7	Rebecca	Kanagaki	Normal	0.632317027492696
8	Kim	Brunner	Bronze	0.777550694555325
9	Brenda	Blumberg	Normal	0.630048566504778
10	Darren	Stanz	Bronze	0.777550694555325
11	Jonathan	Murraiin	Bronze	0.777550694555325
12	Jewel	Creek	Bronze	0.777550694555325
13	Peggy	Medina	Bronze	0.777550694555325
14	Bryan	Rutledge	Bronze	0.777550694555325
15	Walter	Cavestany	Bronze	0.777550694555325
16	Peggy	Planck	Bronze	0.777550694555325



Data Mining Models

Model Name	Algorithm Name	Description
Customer	Microsoft_Decision_Trees	
CustomerNB	Microsoft_Naive_Bayes	
CustomerNN	Microsoft_Neural_Network	
CustomerSegments	Microsoft_Clustering	
Forecast	Microsoft_Time_Series	



演示

更多信息...

- DM URL
  - www.sqlserverdatamining.com
- 新闻组:
  - Microsoft.Public.SQLserver.Datamining
- 更多:
  - KDNugget.com
  - www.MineSage.com - 迈思奇


Thank You!