**Business Intelligence**

**DEIS - Dptº Engenharia Informática e de Sistemas**

**ISEC - Instituto Superior de Engenharia de Coimbra**

Capítulos 12 e seguintes de Delivering BI with Microsoft SQL Server 2005 (Larson, Brian)

Base de Dados: Maxmin (disponível no site do livro)

Apoio complementar: Foundations of SQL Server 2005 BI (Langit, Lynn)

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**I - ALGORITMOS DE DATA-MINING NO SS BI STUDIO**

O SS BI Studio implementa os seguintes algoritmos de data-mining:

1. Classification:
   1. Decision Trees
   2. Naïve Bayes
   3. Neural Networks
2. Microsoft Association (do tipo Apriori)
3. Clustering:
   1. Microsoft Clustering (do tipo aglomerativo)
   2. Microsoft Sequence Clustering (da Microsoft Research, detecta padrões de acções, p.e. sequências de links clicados)
4. Regression: Microsoft Time Series (da Microsoft Research, permite a previsão de valores futuros, baseando-se em “*autoregression trees*”)

A aplicação dos algoritmos aos datasets passa pela identificação da função das diversas colunas (campos), tais como:

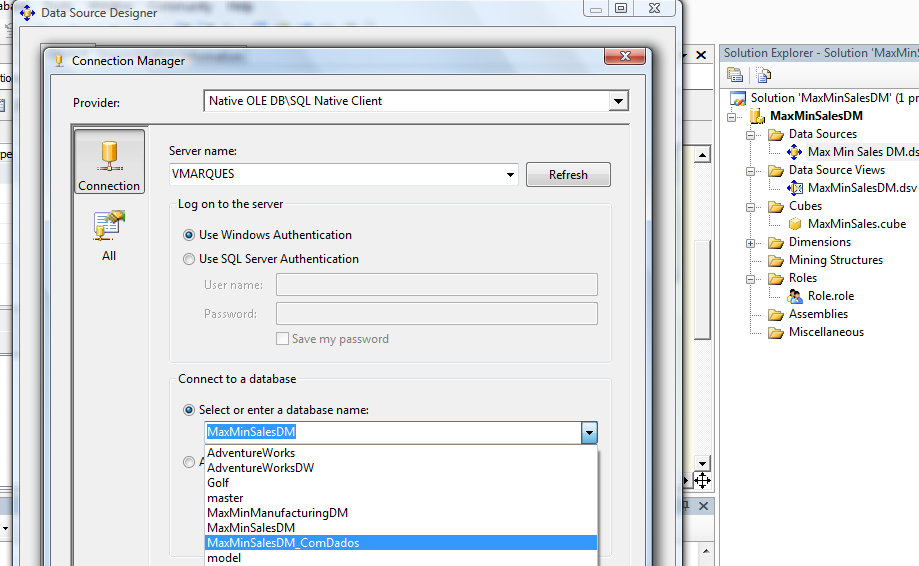
1. ***Key:*** *The key is the unique identifier for a table or a dimension. The key is not used to uniquely identify records or members and it is not used by the data mining algorithm to predict values.*
2. ***Input:*** *Input columns are used by the data mining algorithm when making a prediction. The input values are the values that influence the prediction made by the algorithm.*
3. ***Predict:*** *A predict is a data column whose value is being predicted by the data mining algorithm. This column can also be used as an input column. If data column A and data column B are both predict, then data column A will serve as an input when the algorithm is predicting data column B, and vice versa.*
4. ***Predict Only:*** *A predict only is a data column whose value is being predicted by the data mining algorithm. This column cannot be used as an input column when predicting other columns.*
5. ***Ignore:*** *This data column is not used by the data mining algorithm.*
6. ***Cube Slice:*** *The cube slice enables us to specify an expression that divides the cube into two parts. The portion of the data that satisfies the expression is fed into the data mining algorithm for training. The rest of the data will be testing data set.*

**II - APLICAÇÃO DE UM ALGORITMO DE DATA-MINING**

**Database: MaxMinSalesDM**

**1. Preparação**

1. Entrar no Management Studio, criar a base de dados MaxMinSalesDM\_ComDados.
2. Descompactar o backup da base de dados MaxMinSalesDM e repô-la na base de dados agora criada **ou** fazer o Import do ficheiro Excel Customer.xls para uma tabela ***customers***, consoante o que for fornecido. Se fez importação da tabela a partir de Excel, tem de marcar agora manualmente, na opção Design da tabela, a sua chave primária no campo PK\_Customer\_Name;
3. Entrar no BI Studio e criar o projecto MaxMinSalesDM;
4. Criar uma DataSource para a base de dados criada em 2;



1. Criar uma datasourceview composta apenas pela tabela ***customers*** criada em 2.

**2. Data-Mining: Construção de Modelos de Classificação e Clustering**

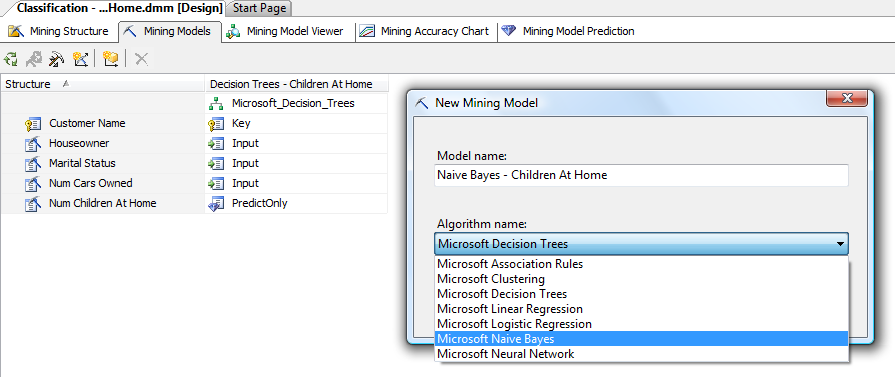
Problema:

*The Maximum Miniatures Marketing Department wants to do a targeted mailing promoting the Mythic World line of figurines. Previous research has shown that the buyers of the Mythic World line of products do not have any children living at home. Unfortunately, the list of people (potencial costumers) does not include the number of children living at home, but it does include the following facts about each:*

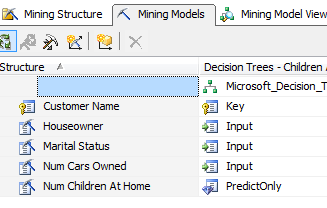
* *Number of cars owned*
* *Marital status*
* *Whether the address is a rental property or occupied by the owner*

*The Marketing Department would like to find a way, using these 3 facts and the information known about current customers (that does have the information about the number of children living at home), to predict which people possibly have more children living at home then the registered value. The mailing will then be sent only to those customers as they are potential customers for Christmas gifts.*

1. Trata-se de uma tarefa de classificação. Algoritmos disponíveis: Decision Tree, Naive Bayes, Neural Network. Vamos usar todos eles (e ainda um de clustering).
2. No Solution Explorer, criar uma nova Mining Structure. Seleccionar “from existing databse or datawarehouse”. *(NOTA: Data Mining Structures can use both relational databases and OLAP cubes for data. We could just as easily have used the MaxMinSalesDM relational database as the source for this cube. Relational databases designed for transactional operations, rather than for a data mart, can also be used)*
3. Escolha, na datasourceview, a tabela *customers* (que fornecerá os exemplos de treino para os modelos)
4. Escolha o modelo Decision Tree. Clique Next e siga o wizard:
   1. Escolha como Predict a coluna Num\_Children\_At\_Home.
   2. A chave primária deverá ser proposta automaticamente como PK\_Customer\_Name. Se não estiver marcada, faça-o manualmente
   3. Clique no botão Suggest. Compreenda as sugestões de colunas do input dadas pelo sistema. Escolha como inputs apenas: num\_cars owned, house-owner e marital\_status;
5. Altere o tipo dos atributos de modo a que sejam todos “discrete” (isto garante que serão tratados como inteiros e não valores contínuos)
6. No nome da estrutura escrever DecisionTrees - Children-at-Home; no nome do modelo escrever Decision Trees - Children At Home. Clique Finish.
7. Aparece a estrutura que tem vindo a desenhar. Abra o separador Mining Models. Para criar outro modelo, basta clicar no New Mining Model. Escolha Naive Bayes com o nome Naive Bayes - Children at Home.



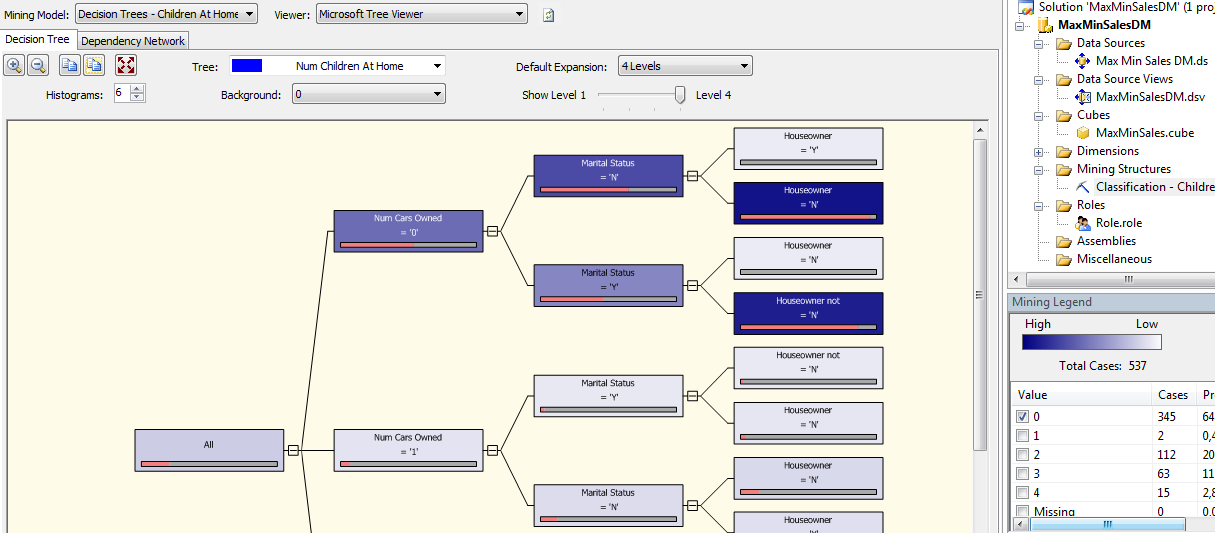
1. Repita a operação anterior (12) para criar os modelos Neural Network, Clustering e Association Rules (use os nomes apropriados em cada caso)
2. Faça Save All
3. Clique Process the Mining Structure and All Its Related Models button na toolbar Mining Models da Design tab toolbar. Clique Run e aguarde o processamento.

****

**3. Data-Mining: Visualização dos Resultados**

**Decision Tree**

1. **Ver a árvore**: Clique no separador Mining Model Viewer para visualizar a árvore:
   1. Há ajustes de zoom, etc.
   2. A caixa Mining Legend mostra as cores associadas a cada target (i.e. o número de filhos residentes);
   3. Pode desmarcar algumas targets e observar apenas a árvore resultante para 1 filho, 0 filhos, etc.
   4. Em cada nó da árvore existe uma barra colorida em que cada cor é proporcional ao número de records de cada target;
   5. As cores de fundo dos nós (rectângulos) variam em função do número de records que contêm. Na dropbox Background seleccione o target que pretende visualizar na cor de fundo (p.e. 0 filhos, o alvo que nos interessa);
   6. Na legenda pode também seleccionar “0”. Há 345 records. A barra rosa é maior em Home Owner=N, MaritalStatus=N, NumberCarsOwned=0, significando que estes valores dos atributos determinam 0 filhos;
   7. Colocando o rato sobre os nós, obtém informação acerca do número de elementos em cada um.



* 1. A árvore “lê-se” da esquerda para a direita. Neste exemplo, a primeira ramificação dá-se no nó “All”, pelo atributo NumCarsOwned, seguindo-se MaritalStatus e finalmente HomeOwner. Simultaneamente os nós vão ficando mais escuros, significando uma MAIOR densidade do target NumFilhos=0. Em concreto, os valores NumCarsOwned=0 + MaritalStaus=N + HomeOwner=N determinam a maior % de NumFilhos=0; outro caminho alternativo será por NumCarsOwned=0 + MaritalStaus=Y + HomeOwner=N, embora mais “fraco” dada a cor mais leve do nó MaritalStatus=Y.

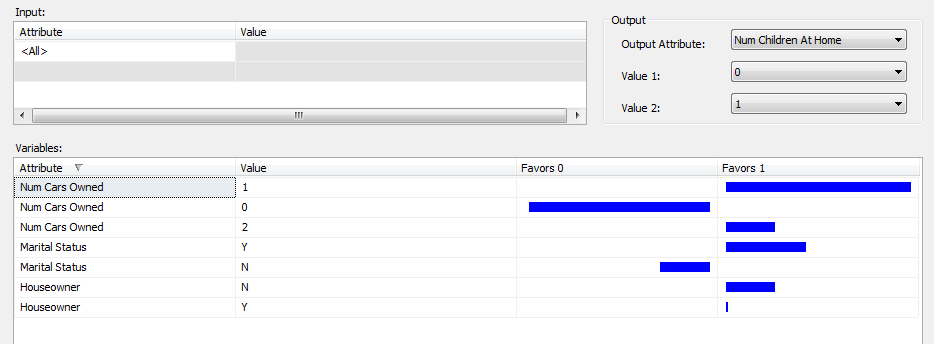
1. **Ver as dependências**: Clique no separador Dependency Network. Clique depois em cada um dos nós do diagrama, em especial no target Num Children at Home. A legenda esclarece o significado das cores de cada nó. O slider do lado esquerdo permite excluir os determinantes mais fracos, o que claramente indica que Num Cars Owned é o atributo que mais fortemente prediz o número de filhos em casa.
2. **Ver os detalhes de conteúdo:** Os detalhes podem ser vistos numericamente seleccionando na dropbox Viewer o modelo Microsoft Mining Content Viewer.

**Naive Bayes**

1. **Ver as dependências e restantes informações:** Clique no separador Mining Model Viewer para visualizar o diagrama de dependência resultante. É análogo ao utilizado nas árvores. Observe os restantes separadores Attribute Profiles, Characteristics e Descrimination (self-explanatory)
2. **Ver os detalhes de conteúdo:** Os detalhes podem ser vistos numericamente seleccionando na dropbox Viewer o modelo Microsoft Mining Content Viewer.

**Rede Neuronal**

1. **Ver determinantes:** Clique no separador Mining Model Viewer para visualizar as dependências do target em relação aos atributos. A rede funciona como um classificador binário escolhendo entre 2 alvos possíveis, seleccionáveis nas dropboxes do lado direito. Por exemplo, a discriminação entre 0 ou 1 filho vivendo em casa, mostra-se na figura seguinte:



**Clustering**

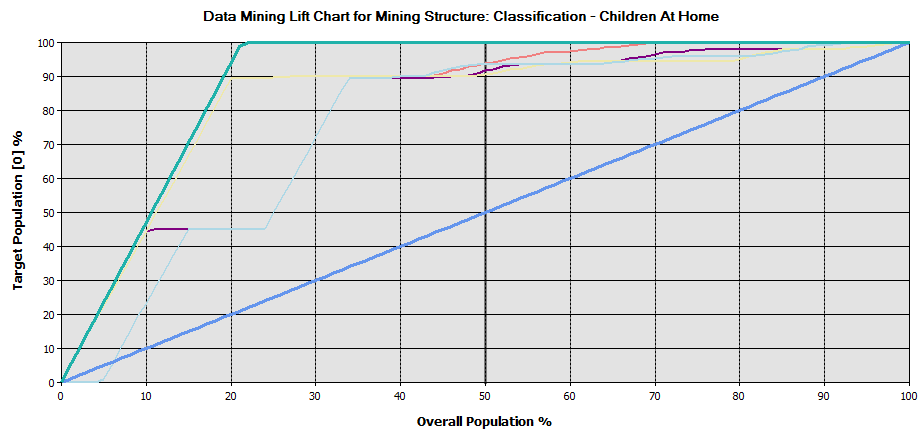
1. **Ver clusters:** Seleccionando clustering, verifica-se que o sistema gerou 8 clusters (o utilizador não tem controlo sobre este número…). Os links entre clusters mostram o grau em que estão relacionados (o slider à esquerda possibilita ajuste de visualização). Em shading variable pode escolher a variável visualizada na cor do clusters e seus links. Escolha Num Children at Home e na dropbox seguinte “0” para aplicar ao caso que pretendemos (i.e. 0 filhos em casa). Note que o cluster 5 é o que tem maior densidade de 0 filhos em casa. Pode renomeá-lo para “No Children at Home” clicando com botão direito;
2. Os restantes separadores são self-explanatory.

**4. Data-Mining: Validação de Modelos; Lift Charts e ROI**

1. Vamos validar o modelo (e ver a sua taxa de erro) usando como TestSet os 30% de clientes que foram reservados quando do treino do modelo (TestCases).
2. Seleccionar o separador Mining Accuracy Chart e o separador Input Selection. Em Select Dataset to Be Used … escolha Use Mining Model Test Cases.
3. Veja o Lift Chart resultante e interprete-o
4. Volte ao separador input Selection.
5. Em Select Predictable…Lift Chart, seleccine 0 (significando que o alvo para o Lift Chart é a classificação 0 filhos em casa). Volte a visualizar o Lift chart e interprete-o
6. Repita 5 para 1, 2, 3 e 4 filhos

***Lift Charts:*** *Um Lift Chart é um gráfico muito utilizado em marketing. Para o construir, ordena-se um TestSet pela probabilidade de cada item ser classificado como positivo. Traça-se então um gráfico cujas abcissas são a % de casos do testset e as ordenadas a % de casos classificados como positivos considerando que os exemplos são aplicados ao classificador pela ordem acima descrita. Ora, nestas condições, se p.e. o testset tiver 100 casos dos quais 20 são positivos, ordenando o testset pela ordem descrita, o gráfico atingirá o valor 100% para 20 casos, e depois manterá uma linha horizontal no valor 1, até que seja considerada 100% da população aplicada ao classificador. Esta curva é a “curva do classificador perfeito” de um Lift Chart.*

NOTA: No diagrama do Lift Chart, a legenda esclarece o classificador a que cada curva pertence. A figura mostra o caso da predição para 0 filhos. O classificador perfeito corresponde à linha verde: como se pode ver, há cerca de 22 casos com 0 filhos. O classificador melhor é a Decision Tree (96%) seguido da Rede Neuronal (94%);

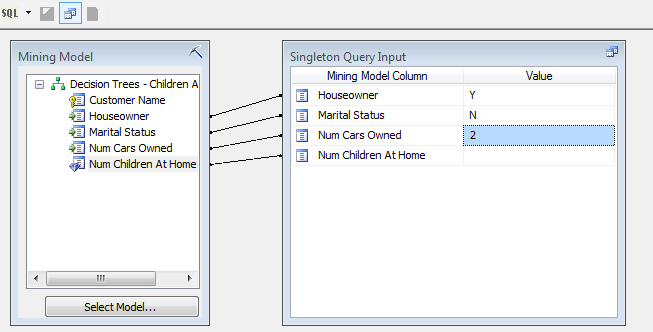


**Classification Matrix**

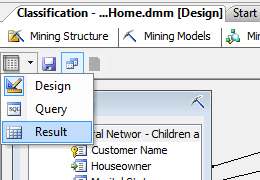
1. O separador Classification Matrix mostra as tabelas de confusão para cada um dos classificadores.

**5. Data-Mining: Aplicação dos Modelos ao Negócio**

1. Vamos agora aplicar casos novos ao modelo;
2. Seleccione o separador Mining Model Prediction. Em Select Model escolha por exemplo a Rede Neuronal.
3. Para testar um só caso (processo 1):
   1. Active o Singleton Query e preencha p.e. conforme figura seguinte;



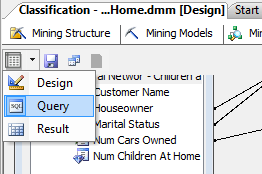
* 1. Para ver o resultado da previsão clique seleccione Query Result conforme figura seguinte. O resultado para este caso será “4 filhos”.



1. Para testar 1 só caso (processo 2):
   1. Desmarque as opções do Singleton Query feito anteriormente
   2. Na janela abaixo, em Source escolha Neural Network
   3. Na 2ª linha, em Source escolha Custom Expression, em Field digite 'Houseowner=Y; Marital Status=N; Num Cars Owned=2' e em Alias (opcional) escreva p.e. InputCriteria
   4. Mude a vista para Query Results. A previsão para estes casos será de “2 filhos”
2. Para testar N casos:
   1. Clique de novo no botão Singleton Query para desactivar este modo;
   2. Escolha o modelo Neural Network;
   3. Em Select Case Table escolha a tabela Customers
   4. Em baixo indique os campos que pretende visualizar: da tabela Customers, os campos PK\_Customers\_Name (e pode indicar outros, como a morada, por exemplo) e Num-Children\_At\_Home. Da Neural Network, o campo Num\_Childre\_At\_Home também. Assim poderemos comparar o valor na tabela com o resultado predito pelo classificador;
   5. Mude para Query Results. A previsão do número de filhos para cada cliente aparece;
   6. Por exemplo para ver apenas os que têm 0 filhos (lista dos potenciais clientes pretendidos), voltar a Design view e na 1ª linha, em Criteria/Argument escrever=0; voltar a Query Results: agora aparece apenas a lista de pessoas com previsível número de filhos =0;
   7. Por exemplo para ver apenas por exemplo clientes que tenham mais filhos previstos do que o número registado, colocar em Criteria a expressão “>Num\_Children\_At\_Home”.
3. Para salvar o resultado de modo a ser utilizado na correspondência a enviar:
   1. Clique em Save Query Result
   2. No nome da tabela coloque Max Min Mailling
   3. Uma tabela com este nome é criada na DW e preenchida com o resultado do query. Pode vê-la entrando no SS Management Studio.

**6. Data-Mining: Queries em DMX** (Data-Mining Extensions)

1. Ao preencher os formulários de design na secção anterior, o BI Studio gerou um query na linguagem DMX (Data-Mining Extendions) disponibilizada pelo SS BI Studio. Pode vê-lo clicando em Query:



SELECT

[Neural Network - Children at Home].[Num Children At Home],

t.[Fname],

t.[Lname],

t.[Address1\_Name],

t.[City\_Name]

From

[Neural Network - Children at Home]

PREDICTION JOIN

OPENQUERY([Max Min Sales DM],

'SELECT

[Fname],

[Lname],

[Address1\_Name],

[City\_Name],

[Num\_Cars\_Owned],

[Houseowner],

[Marital\_Status]

FROM

[MaxMinSalesDM].[Customer]

') AS t

ON

[Neural Network - Children at Home].[Num Cars Owned] = t.[Num\_Cars\_Owned] AND

[Neural Network - Children at Home].[Houseowner] = t.[Houseowner] AND

[Neural Networ - Children at Home].[Marital Status] = t.[Marital\_Status]

WHERE

[Neural Network - Children at Home].[Num Children At Home] =0

1. A sintaxe do DMX pode ser estudada no livro guia. Contudo a sua utilização é pouco frequente dada a interactividade proporcionada pelo modo de Design (o que foi utilizado atrás)