Taking Your Application Design to the Next Level with Data Mining

Peter Myers

Mentor – SolidQ Australia

HDNUG – 24 June, 2008

Copyright © 2008, Solid Quality Mentors. All rights reserved.

WHO WE ARE

- ***Industry experts:***

  Growing, elite group of over 90 of the world's best technical experts who, as reflected by the high concentration of Microsoft MVP's and RD's in our ranks, achieve excellence in their industry by maintaining the highest credentials.
- ***Published authors:***
  Best technical reference books, Microsoft reference materials, industry white papers, technical magazine articles, and webcasts.
- ***Top technical speakers:***
  PASS Community Summit, Microsoft TechEd, The Microsoft BI Conference, SQL Server DevConnections, countless user groups, international conferences and events.
- **For more information visit *www.solidq.com.au***

Copyright © 2008, Solid Quality Mentors. All rights reserved.

WHAT WE DO

Provide advanced, world-class expertise across the entire Microsoft relational data and development platforms and complimenting technologies.

PRACTICE AREAS		SERVICES
Relational Database Management		Advanced, Public Training
Business Intelligence		Customized, Private Training
Development Methodologies		Solution Delivery & Tuning
SharePoint Collaboration		Enhanced, Mentoring Services

For more information visit *www.solidq.com.au*

Copyright © 2008, Solid Quality Mentors. All rights reserved.

AGENDA

- Introducing Data Mining
- SQL Server™ 2008 Data Mining
- Data Preparation
- Describing the Data Mining Process
- Data Mining Visualization
- Demonstrations

Copyright © 2008, Solid Quality Mentors. All rights reserved.

INTRODUCING DATA MINING

- Addresses the problem:
  "Too much data and not enough information"
- Enables data exploration, pattern discovery, and pattern prediction—which lead to knowledge discovery
- Forms a key part of a BI solution

Copyright © 2008, Solid Quality Mentors. All rights reserved.

BUSINESS SCENARIOS

- Identifying responsive customers/unresponsive customers (also known as churn analysis)
- Detecting fraud
- Targeting promotions
- Forecasting sales
- Cross-selling

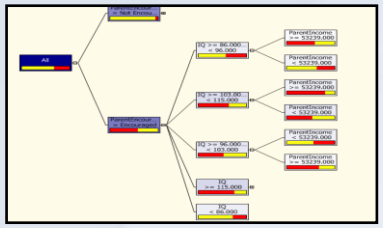
Copyright © 2008, Solid Quality Mentors. All rights reserved.

SQL SERVER™ 2008 DATA MINING

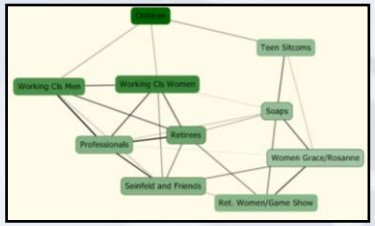
- Hides the complexity of an advanced technology
- Includes full suite of algorithms to automatically extract information from data
- Handles large volumes of data and complex data
- Data can be sourced from relational and OLAP databases
- Uses standard programming interfaces:
  - XMLA
  - DMX
- Delivers a complete framework for building and deploying intelligent applications

Copyright © 2008, Solid Quality Mentors. All rights reserved.

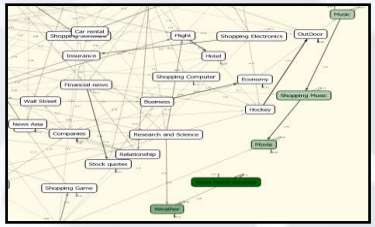
SQL SERVER™ 2008 ALGORITHMS



- Decision Trees
  - The most popular data mining technique
  - Used for classification



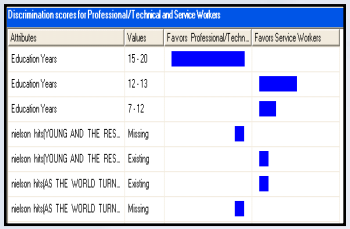
- Clustering
  - Finds natural groupings inside data



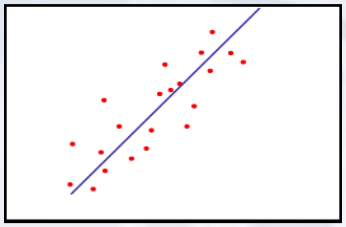
- Sequence Clustering
  - Groups a sequence of discrete events into natural groups based on similarity
  - Use this algorithm to understand how visitors use your Web site

Copyright © 2008, Solid Quality Mentors. All rights reserved.

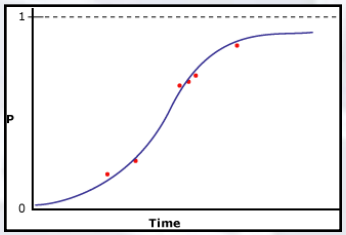
SQL SERVER™ 2008 ALGORITHMS



- Naïve Bayes
  - Used for classification in similar scenarios to Decision Trees



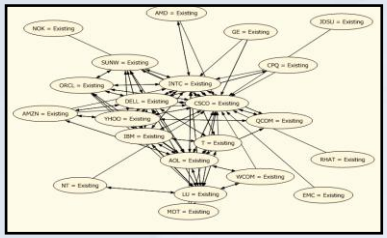
- Linear Regression
  - Finds the best possible straight line through a series of points
  - Used for prediction analysis



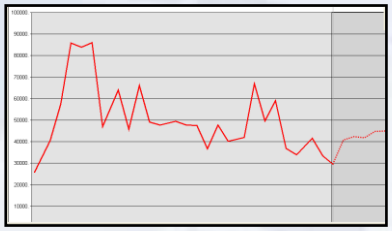
- Logistic Regression
  - Fits to an exponential factor
  - Used for prediction analysis

Copyright © 2008, Solid Quality Mentors. All rights reserved.

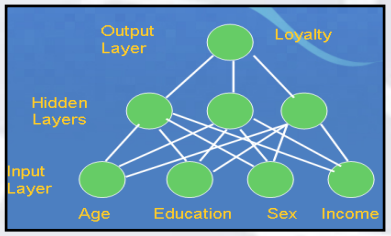
SQL SERVER™ 2008 ALGORITHMS



- Association Rules
  - Supports market basket analysis to learn what products are purchased together



- Time Series
  - Forecasting algorithm used to predict future values from a time series
  - Has been improved in SQL Server 2008 to produce more accurate long-term forecasts



- Neural Net
  - Used for classification and regression tasks
  - More sophisticated than Decision Trees and Naïve Bayes, this algorithm can explore extremely complex scenarios

Copyright © 2008, Solid Quality Mentors. All rights reserved.

DATA PREPARATION

- Often significant amounts of effort are required to prepare data for mining:
  - Transforming for cleaning and reformatting
  - Isolating and flagging abnormal data
  - Appropriately substituting missing values
  - Discretizing continuous values into ranges
  - Normalizing values between 0 and 1
- Of course, having the required data to begin with is important:
  - When designing systems, give consideration to attributes that may be required as inputs for classification
    - For example, demographic data: Age, Gender, Region, etc

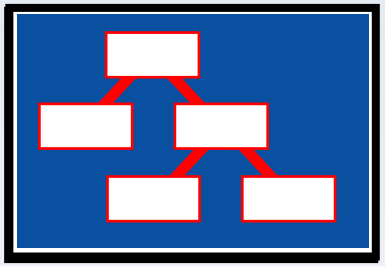
Copyright © 2008, Solid Quality Mentors. All rights reserved.

DESCRIBING THE DATA MINING PROCESS

Design time

Process time

Query time

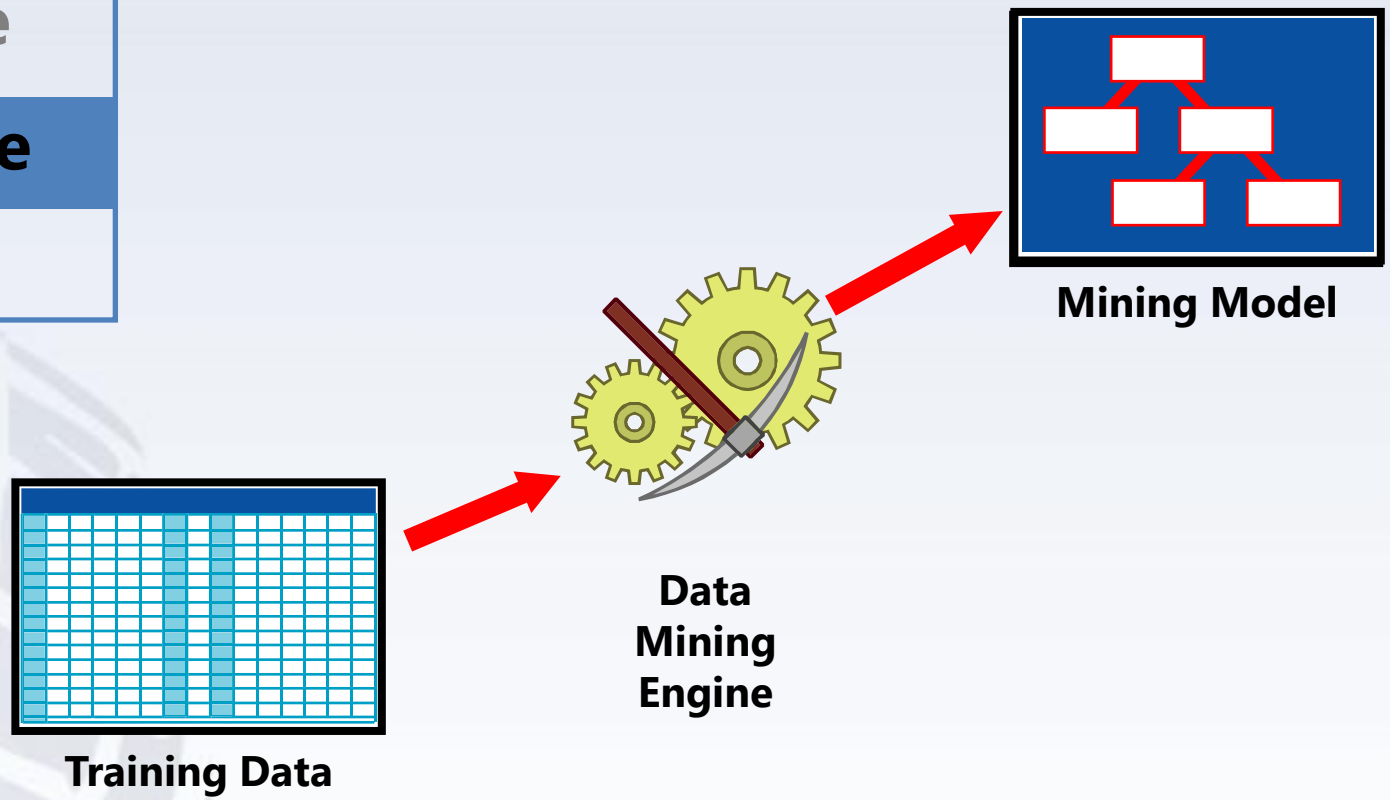


Mining Model

Copyright © 2008, Solid Quality Mentors. All rights reserved.

DESCRIBING THE DATA MINING PROCESS

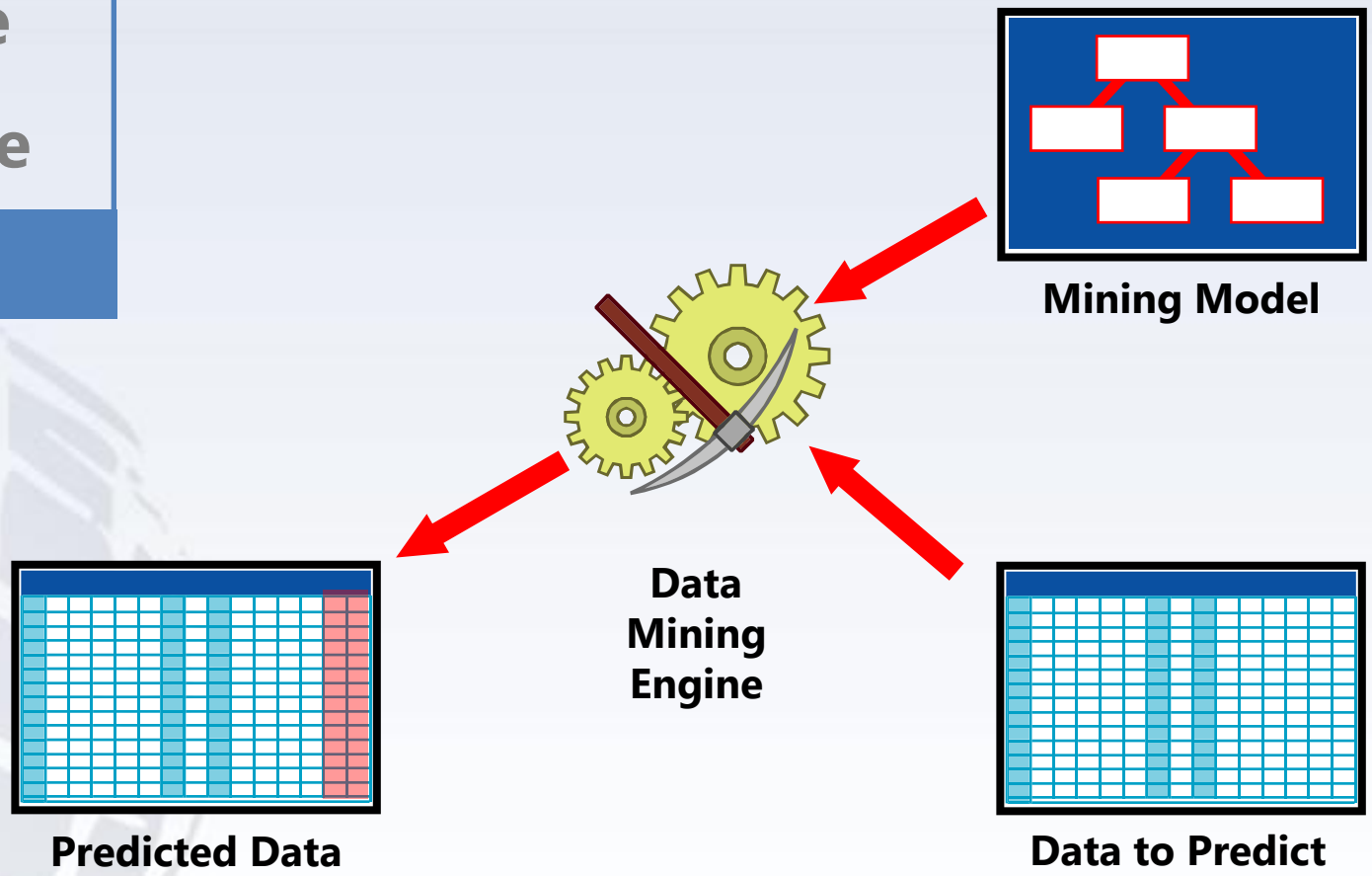
- Design time
- **Process time**
- Query time



Copyright © 2008, Solid Quality Mentors. All rights reserved.

DESCRIBING THE DATA MINING PROCESS

- Design time
- Process time
- **Query time**



Copyright © 2008, Solid Quality Mentors. All rights reserved.

DATA MINING VISUALIZATION

- In contrast to OLTP and OLAP queries, data mining queries typically extract information that the user is not aware of
- Appreciate that end users do not typically query data mining models directly
- Visualizations can effectively present data discoveries
- SQL Server™ 2008 provides algorithm-specific visualizations that can:
  - Test and explore models in BIDS
  - Be embedded into Web and Windows Forms applications
- Developers can construct and plug-in custom data mining viewers

Copyright © 2008, Solid Quality Mentors. All rights reserved.

DEMONSTRATIONS

1. Developing a Mining Structure
2. Embedding a Data Mining Report
3. Embedding a Data Mining Visualization
4. Programming Automatic Data Validation
5. Enhancing an E-Commerce Site with Market Basket Analysis

Copyright © 2008, Solid Quality Mentors. All rights reserved.

RESOURCES

- www.microsoft.com/sql/technologies/dm
  - Links to technical resources, case studies, news, and reviews
- www.sqlserverdatamining.com
  - Site designed and maintained by the SQL Server Data Mining team
  - Includes: Live samples, tutorials, webcasts, tips and tricks, and FAQ
- Data Mining for SQL Server 2005, by ZhaoHui Tang and Jamie MacLennan

Copyright © 2008, Solid Quality Mentors. All rights reserved.