


Data mining in the "real world"

What do we need & what do we have

Françoise Soulié Fogelman
KXEN
Francoise@kxen.com

KDD 2006

Data Mining for Business Applications Workshop
Philadelphia, USA
August 20 - 23, 2006



Agenda

- Data mining in companies today
- Requirements for extreme data mining
- Extreme data mining examples
- What is still missing

2

Data mining in companies ... is widely used

- **Customer Relationship Management**
 - Customer knowledge
 - Propensity scoring
 - Churn, cross-sell, up-sell
 - Life-time value
 - Campaign optimization
- **Risk management**
 - Credit scoring
 - Fraud detection
- **Web**
 - Profiling

... somewhat in ...

- **Production**
 - Defect & quality analysis
- **Pharmacy**
 - Drug design ...

... but is often seen as « an art » of the expert ... thus not ubiquitous

Mainly in :

- Telco
- Bank, insurance
- Retail
- E-business

3

Companies have data ...

- Companies have created customer datawarehouses
 - Size up to 100s Tb and increasing fast (X3 every 2 years)
 - Investment up to \$ 100s M
 - Millions of customers & Thousands of variables

... they want to see ROI

DataBase Size (TB)

Nb of rows, records or objects (Millions)

From <http://www.wintercorp.com>

4

... but do companies exploit their data ?

- **Production of models**
 - Relies - still - solely on Analytics Department
 - Serves the needs of 20-50 business users
 - Produces 5-10 models / year
 - ... with 5-10 DM experts who ... are scarce
 - Involves a triangular relationship with IT and business users
- ... thus requires a typical 3-8 weeks delivery time
- ... thus fails to optimize many business processes


5

Agenda

- Data mining in companies today
- Requirements for extreme data mining
- Extreme data mining examples
- What is still missing

6

Extreme data mining




- **Top-tier companies want to improve all their processes**
 - which could require 1000's of models
- **... handling terabytes of data**
 - Millions of customers
 - Thousands of variables
- **... with limited resources**
 - 5 DM experts for 1000 models per year
 - Sometimes no DM expert at all !

The only possible answer is productivity through automation and de-skilling providing companies with the ability to deploy « model factories »

© VVENI, Confidential 7

Extreme data mining




- **Handle multiple data sources**
 - Increasing volumes
 - Heterogeneous sources
- **Produce models**
 - Fast, accurately
 - In real-time
- **Fit multiple users' needs**
 - Business users need it
 - They want to do it themselves
 - ... without having to become data mining experts
- **Integrate into the Information System**
 - Read any data format
 - Write to any format
 - ... to be used within global applications
- **Deliver value**
 - Accurate, actionable models
 - Productivity

- Avoid duplicating data
- Model calibration ... fast
- Model application ... fast
- Can be used by business users
- Relies on a sound methodology
- Data mining standards compliant
- Include automatic control
- Return On Investment

© VVENI, Confidential 8

Handle volumes




- **Data base includes**
 - Millions of rows (e.g. customers)
 - Thousands of variables
 - In various formats and types
 - Continuous, nominal, ordinal,
 - Text, image, speech, ...
- **... of – often – poor quality (missing, outliers, ...)**
- **Handle all this data**
 - Globally
 - Must not look individually at each variable
 - Must not hand-pick « relevant » variables
 - Solving data format issues
 - Must recode all data on-the-fly
 - Accepting poor quality
 - Must process automatically outliers & missing values

Fully automated data handling ... whatever the volume

© VVENI, Confidential 9

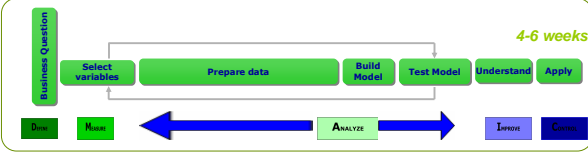
Produce models



- **Model production**
 - « Velocity is always king ! Good enough and deployed is always better than perfect and in the lab ! » Barclays-Teradata Conf – September 2005
 - Data preparation / recoding takes 80 % of the time

Fully automated data prep/recoding & model production

4-6 weeks




- **Real-time**
 - At the time of contact
 - In-bound call, Web

Fully automated model application

© VVENI, Confidential 10

Limited DM expertise




- **Users know their business**
 - What are the key issues & where are the key problems
 - What are the data used / generated by their activity
 - What is the business value of the result delivered by a model
- **Users do not know statistics (and don't care)**
 - What is the best algorithm to be used when
 - How to manipulate data
 - How to select / code variables according to statistical significance
 - Handle : outliers, missing data
 - How to "decode" a model results
 - Model must be self-explaining
 - How to evaluate the statistical validity of a result

De-skill model production & open-up to business users

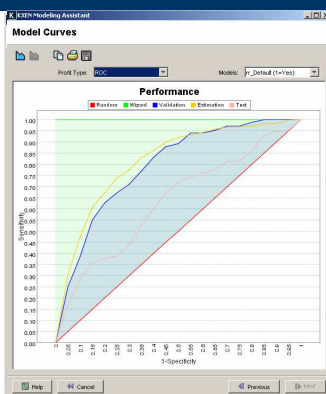
© VVENI, Confidential 11

Deliver value



- **Accuracy**
 - Ability to produce accurate models ... on future data
 - We need a way to assess that ability

Fully automated model quality assessment



© VVENI, Confidential 12

KxEn Deliver value

- **Productivity**
 - Ability to produce many models

De-skill model production

Vodafone needs for Tier 1 Telco

	# Analysis /Year
• Segmentations 2*2*10	40
• Churn in General 2*3*2*3	36
• Churn per product 2*3*2*4*10	480
• Cross sell : segments*offers 2*4*10	80
• Acquisition 2*4*10	80

This means trying to create 716 models per year...

From Vodafone - Teradata Conf - Sept. 2005

© KVENI/Confidential 13

KxEn Agenda

- Data mining in companies today
- Requirements for extreme data mining
- Extreme data mining examples
- What is still missing

© KVENI/Confidential 14

KxEn Have we succeeded ?

Examples

- Number of variables


1. 900
2. 1 000
3. 1 200
4. 2 500
5. 4 200
6. 5 800

How to handle such large dimensions with conventional tools ?

© KVENI/Confidential 15

KxEn Have we succeeded ?

Examples



1.
 - Models implemented in 2 days
 - Cut operational costs 50% and time to model 90%
 - Score 75m households in Teradata in 30 minutes
2.
 - 20 predictive models built & deployed by 2 persons in a month
 - Reduce data preparation from 70% of effort to almost nothing
 - Automatically maintain hundreds of models a year
 - Expect 10x productivity gain
3.
 - 1680 production models/year
4.
 - "I built 377 models with 100K rows each in a couple hours on only a PC"
 - 10 models in 3 days
5.
 - 800 seconds on 1,000,000 observations with 220 variables

© KVENI/Confidential 16

KxEn Agenda

- Data mining in companies today
- Requirements for extreme data mining
- Extreme data mining examples
- What is still missing

© KVENI/Confidential 17

KxEn Missing features

- Ability to handle / code multimedia formats
 - Text
 - Images
 - Speech, sound ...
- Use of unlabelled data
- Vertical integration of data mining everywhere ...

Meanwhile, let us solve the million "easy" problems around

© KVENI/Confidential 18