Some Data Mining project lessons

April 2004

Frank Meyer – France Telecom

Topics

1. **Some data mining applications in Telecom Industry**
   - Complex (very difficult) applications
   - Usual (may be not so easy) applications
   - Easy applications
2. **Main problems in industrial data mining**
   - Data management
   - Model deployment & maintaining
3. **Data Mining & Machine Learning**
   - Some comparisons…

Conclusion…

France Telecom context

In the world

- 117, 000, 000 clients
- (240,100 employees on 5 continents)

In France

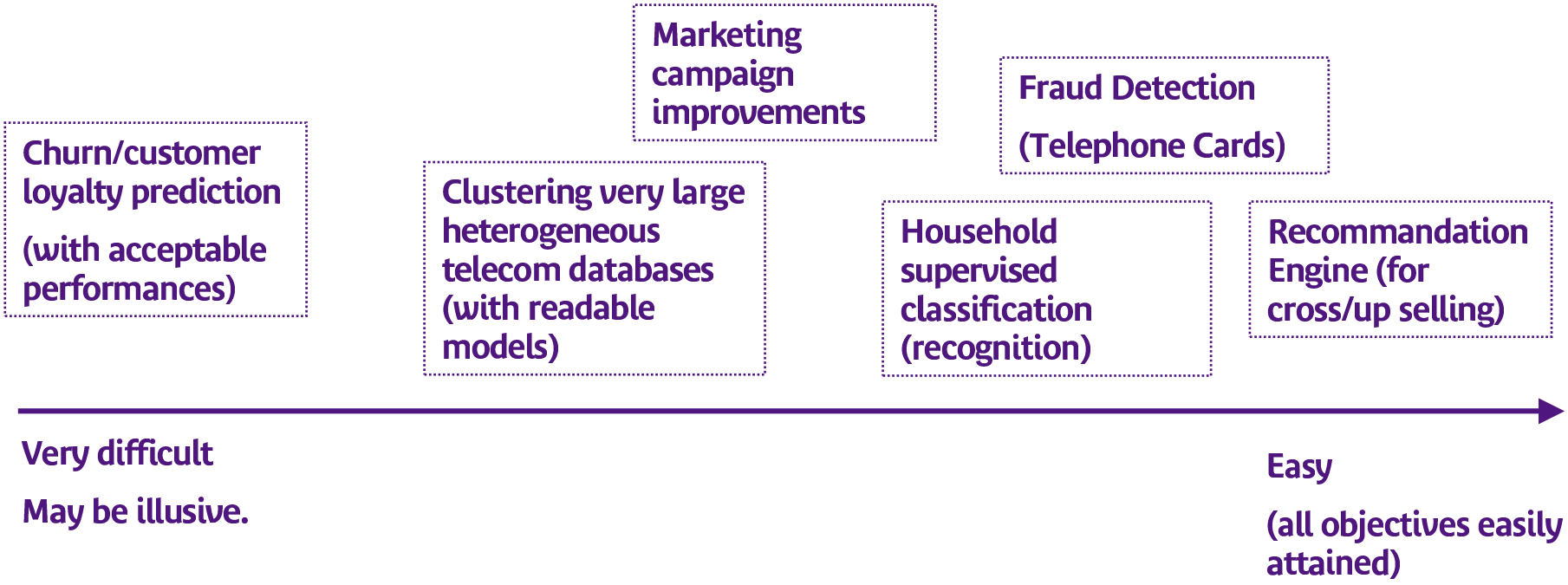
- More than 30, 000, 000 clients
- (120 000 employees)

FT R&D

- More than 3700 employees
- DTL/TIC Lab: 25 (data mining: 10 engineers and researchers full time workers - others 15 also diagnostic and image appl.)

1. Some data mining applications …

Examples of projects from Telecom Industry…



France Telecom experiences

Typical very difficult application

Churn prediction

- Goal: You know that Mr Smith has
  - *a bill about $124.5*
  - *a large increase of his communications every saturday afternoon*
  - *Etc… (a lot of information)*
- And you want to give Mr Smith a no-loyalty probability (risk to go to a competitor)
  - 0%: Mr Smith will certainly stay
  - 100%: Mr Smith will certainly leave our company (in 2 months, 4 months…)
- currently not enough true positive recognition rate. Even with many well chosen features

Typical "not so easy" applications

Clustering large telecoms datasets

- Goal: build synthetic, readable, usefull clusters from
  - Tables with millions of records
  - Tables with hundreds of attributes
  - Tables with mixed-type attributes
  - and: some objects of interest have weak frequencies
  - and: some attributes are irrelevant or very noisy

- Even if we can find clusters
  - Why my a priori strategic segmentation is not better ?
  - How can I understand the result ?
  - Why this group is in this cluster ?

- what is the good algorithm for this task ?

Typical "not so easy" applications

Marketing campaign improvements

- goal: Make use of returns of the previous campaign to find rules to better understand what happened, to build better targets for the next campaign

- Example of classical pitfall: (not a real one, but similar examples exist)
  - A marketer says "I think that a good target for ADSL may be the city of Plouigneau" (little city in Brittany, France)
  - A sell campaign is done on Plouigneau.
  - Then somebody says "What about using data mining high-technology to improve the campaign ?"
  - A data mining analysis is done.
  - The model is very clear and shows high potential target... in Plouigneau.

Typical easy applications

Fraud detection (phone cards)

- Goal: learn to detect true positive fraud only
- stereotyped behavior, easy to detect
- few attributes to consider, each attribute very reliable and meaningfull (example: number of calls / day)

Household (supervised) classification

- goal: learn to classify records of clients
- predefined social categories : Single, Adults with teenagers, Seniors, Adults without child
- some typical seniors, teenagers behavior... can be seen.

Typical easy application

- **Recommandation engine "people who bought X also bought Y"**

- **Key points**
  - Algorithm: the simplest, the best.
  - learning process is based on reliable data: purchases on a website (ID of products, ID of customers)
  - Self-maintaining model
  - Easy deployment

Usual Data Mining Industrial Process

- **For example: CRISP DM Methodology:**
  - www.**crisp-dm**.org

- **6 steps:**

1. Business understanding
2. Data understanding
3. Data preparation
4. Modelling
5. Evaluation
6. Deployment

2. Main problems in industrial Data Mining

- **… probably currently without good solution.**

What are the main problems ?

Data management costs

- Data availability
- Data extraction (select and merge tables)
- Data re-engeneering (often data understanding)
- Data preparation

Model deployment & maintaining

- Deployment and integration into a process
- Rebuild the model when its performances decline

Data Management: a critical issue

- **Data preparation ?**
  - select data, clean data, recode data…
  - build aggregates…
- **Some learning methods (trees/rules based) can be used to reduce costs**
  - no data type restrictive
  - no scale sensitive
  - tolerant to noise
  - => less needs of data preparation
- **Still about 80% of the cost of a data mining project**

Model deployment & maintaining

- **The model is built and evaluated (in a lab)… And after ?**
- **How to deploy and maintain ?**
- **Still a very important problem**
  - Mecanics, Physics… => stable models: same causes, same effects
  - This is not true (because of always changing environment):
    - in marketing
    - even in a lot of industrial processes…
  - So Models have a short life-time (we have to rebuild them)
- **And unfortunately:**
  - Databases and Information Systems are not static
  - Some new data appear, some data become unavailable

3. Machine learning / Ind. Data Mining

- **Comparison of approaches**
- **and important research areas …**

Data Mining & Machine Learning

- **Well known dataset repositories do not represent real-industrial applications today**
  - A small table at FT R&D DM Lab:
    - 8000 instances x 200 attributes
  - UCI repository
    - Iris, Ionosphere… toy datasets
    - Even Letters (20000 instances x 16 attributes) is not really representative
  - Algorithm behaviors on large databases often change.

Data Mining & Machine Learning

you have to understand what the model has learned

- Black boxes are not reliable
  - In complex project, even with a good methodology, you may use bad attributes (taboo attributes: not really available, or using "future knowledge").
  - If the model shows bad performances, and if it is a black box, what can we do ? Where is the problem ?

- What is the most useful: knowledge you cannot share, or knowledge you can share ?

- Data mining is often used in decision making processes

Data Mining & Machine Learning

A good learning algorithm must be scalable

- ... I can always sample: False.
  - weak frequencies for some objects of interest
  - in high dimension, each instance is important
- Error rate performance is the only important issue: False.
  - compromise between several properties
  - the speed of the algorithm is also very important
- Learning algorithms are only used to build a model: False
  - often used to analyse data, to select attributes, to construct aggregates, to test hypothesis
  - often used many times during different tasks in a data mining project (so the quicker, may be the better...)

Data Mining & Machine Learning

- **Is the classical off-line learning principle the best adapted to industrial purposes ?**
  - Incremental on-line algorithms
    - self maintaining
    - rapid
    - => reduce deployment cost
  - Reinforcement learning
    - May be well suited for maketing campaigns
      - *incremental test/evaluation of new targets*
      - *reactive model*
  - More generally data miners need "interactive" methods
    - very quick methods
    - but also readable methods

Conclusion

Many interesting research areas

1. Data mining processes…
2. Comparison of learning algorithms in industrial contexts (not synthetic…)
3. Self-maintaining, rapid, interactive learning systems
4. Learning Systems with multi sources and multi features types integrations abilities

Annex

- **churn prediction performance example**

Precisions on Churn prediction

Churn prediction issue

- If you know that Mr Smith has
  - *a bill about $124.5*
  - *a large increase of his communications every saturday afternoon*
  - *Etc…*
- Do you think you can predict that Mr Smith will go to your competitor in 4 months and 5 days at 3 p.m. ?
- May be… for some churn prediction application editors…

Typical performance

- With a usual 2% churners rate, is a gain factor (lift) of 7 a good performance ?
  - gain factor : 7 (but 7 on 2% churners, not on 13% churners !)
  - maximum true positives detected: 14% (on best scores only)
  - minimum error rate for churner detection : 86% (on best scores only)