Datamining and Its Challenges in a Banking Environment

Chen Wei Xu

Customer Knowledge Management

Bank of Montreal

chen.xu@bmo.com

December 10, 1999

Data Mining

- Massive amounts of data
- Lack of a priori knowledge (especially causal relationships)
- Dynamic environment - ever changing context
- Lack of data (sorry, you just said lots of data…)
- You want the data to tell you something about the customers!

Value in Customer Data

- Banks expect data to answer questions such as:
  - Will the customer buy product X? Do I need to stimulate them to buy (e.g., by mail, with incentive)?
  - Will the customer leave us in the next while? Why? What can I do to stop that?
  - What causes customers to reduce their share of wallet with us? Rates? Service? …?
  - What customer segments do we have? (For marketing, product design, channel management, …)

Types of Data

- Customer data:
  - Age, account types and balances, transactions, open/mature dates, interest rates, postal code, …
- StatsCan - Neighborhood data:
  - Avg income, RRSP contributions, estimated financial assets, age distribution, family structure, dwelling types, …
- Other data:
  - Credit data (delinquency, credit usage, risk score,…)
  - Survey data (hobbies, spending, attitude, …)

Customer Databases

- Millions of records, 1000+ of variables
- Month-end snapshots
- Usually RDBs
- Some history possible (e.g., 12-month), but generally insufficient for time series analysis
- Identity issues
- Stability of data definition

Example 1: Acquisition via Direct Mail

- Context: Acquire new business (products) by direct mail
- Problem: Don’t know who to mail to. Random mailing creates junk mail and costs more
- Solution: Mail only to those likely to generate value
- Benefits: Better ROI, less junk mail

Example 1

- Two models
  - A response model (Who would respond when mailed?)
  - A value model (How much $ if respond?)
- Process
  - Model development
  - Model implementation - target set selection
  - Campaign execution
  - Measurement & validation

Example 1

- Model development
  - Objective: predict mail response/value behavior
  - Define universe
  - Take a random sample (usually n-th sample)
  - Run a pilot campaign (i.e., mail randomly)
  - Collect response & value data from campaign
  - Model creation (with held-off data for validation)
    - Regression, black-box, tree-based, and so on
    - P(Response) = f(x1,x2,…,xn)
    - E(value|response) = g(z1,z2,…,zm)

Example 1

- Model Implementation
  - Score the universe with both models
  - Create overall score S = P(resp.) x E(value|resp.)
    - S := Expected value when mailed
    - S_0 := Expected value when not mailed (constant)
    - $\Delta S := S - S_0$, or incremental value from mailing
  - Rank order customer list by ΔS
  - Cut off list at the customer where
    - Marginal profit (value-cost) becomes zero, or
    - Your maximum budget allows but marginal profit >0.

Example 1

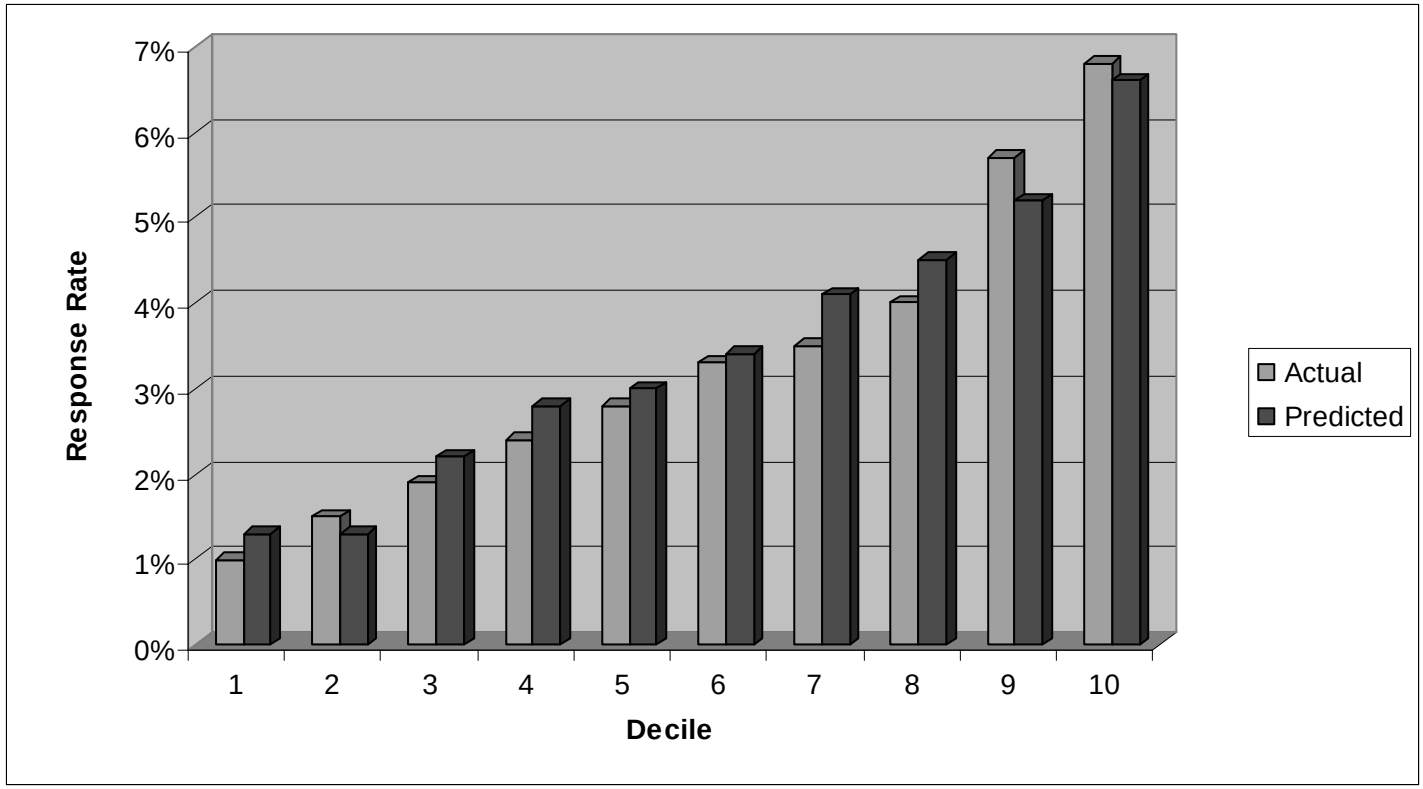
- Execution: Target customers mailed, with a random subset held off as control (no mail) ...

Example 1

- Campaign Measurement - Was it a success?
  - Each target mailing generated value *a* and cost *c*
  - Each control subject generated value *b*
  - Total value generated from campaign is
    - Total Value = N(*a*-*b*-*c*), where N = # of mailings
- Challenges
    - Should *a*, *b* include halo effect?
    - What is the life span of *a*, *b* ?

Example 1

- Model validation
  - Compare prediction and actual
  - Do it every year, for models “wear out”

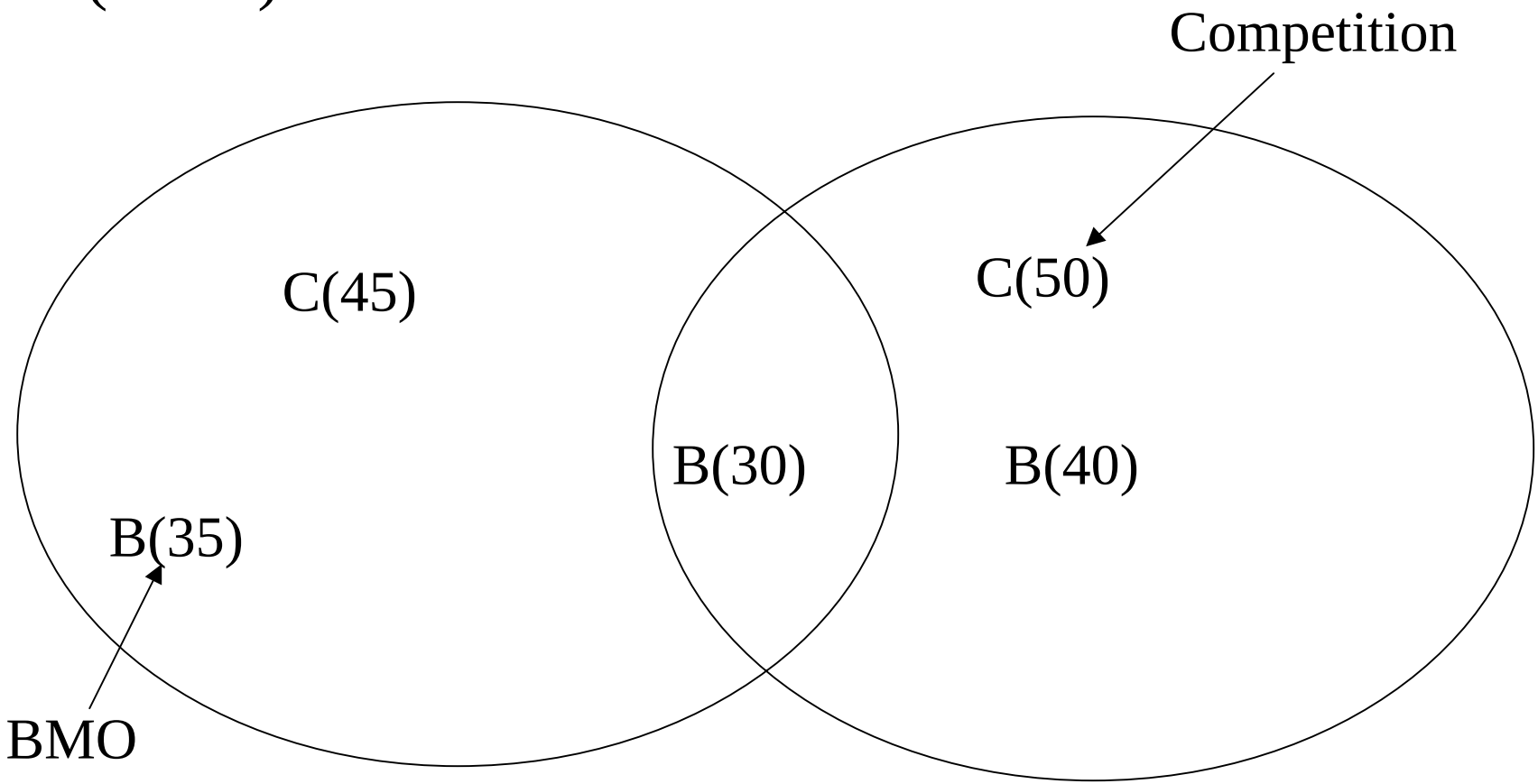


Example 2: Model Fit

- Problem
  - Have n observations (Yi, X1i, …,Xpi), i=1,…n; p large
  - Need a model (e.g., a value prediction model)
    - Y = f(X), where X is a subset of {X1,…,Xp}
  - Q1: How to isolate X?
    - Correlation - captures linear relation (&simple nonlinear)
    - No existing knowledge about causal relations
    - Other methods (such as discrimination analysis) may not work well with continuous Y variable
  - Q2: How to determine f(.)?
    - Usually use linear approximation, but not always works
    - Infinite possibilities of nonlinear function forms
    - Polynomial - order and terms?

Example 3: Distribution

- Branches to match competition service level (hours)



Example 3: Distribution

- Objective: Minimize cost of increasing branch hours (cost is a function of incremental time and branch size)
- Constraint: Match all competitors' service hours for each service area
- Can be formulated as an integer (0-1) programming problem
- Have a greedy heuristic algorithm developed (proven to be suboptimal)