Data Mining Opportunities in Health Insurance

Methods Innovations and Case Studies
Dan Steinberg, Ph.D.

September, 2008

Copyright © Salford Systems 2008
Analytical Challenges for Health Insurance

• Competitive pressures in marketplace make it imperative that insurers gain deep understanding of business
• Essential to leverage the insights that can be extracted from ever growing databases (including web interaction)
• Rich extensive data in large volume allow detailed and effective analysis of every aspect of business
• Areas amenable to high quality analysis include
  – Risk: Probability of Claim, Expected Losses on claims
  – Fraud: Identification of probable individual fraud, detection of organized professional fraud
  – Analytical CRM: precision targeted marketing, scoring policy holders for lapse probability, identifying upsell opportunities
Analytical Opportunities

• “Have Data Will Analyze”
  – A predictive enterprise applies analytical modeling techniques to all areas of business
  – All you need is adequate historical data

• Analytics can be applied in nontraditional ways
  – What makes 2007 different from 2006?
  – Which case managers are most effective for specific types of claim?
  – When is the best time to make a cross-sell offer?

• Opportunities are limited only by creativity of analysts
  – Ad-hoc queries can be reformulated as mini-data mining projects
Why Data Mining Has Changed the Game

• Conventional statistical models (GLMs) take too long to develop and require too much expertise
  – Not enough statisticians to develop all needed stats models
  – Data mining models can be built in far less time

• Data mining has raised the bar for the accuracy that can be achieved
  – Modern methods can be substantially better than GLMs

• Data mining methods can also work effectively with larger and more complex data sets
  – Can easily work with hundreds, even thousands of predictors
  – Can rapidly detect complex interactions among many factors
Importance of Interactions

• “In matters of health everything interacts with everything”
  – Quote from a veteran consultant to the health insurance industry

• Conventional statistical models are typically *additive*
  – Each predictive factor acts in isolation
  – E.g. What is protective effect of large doses of Vitamin E for coronary heart disease?
    • Truth appears to be an interaction: for people under 55 years old the benefit is zero; for over 55 it is substantial

• Certain data mining techniques such as CART and TreeNet are specifically designed to find interactions automatically
  – Conventional stats poorly equipped to detect interactions
Further Data Mining Capabilities

• Data mining methods solve data preparation challenges:
  – Automatic handling of missing values. Generally missing values require considerable manual effort by GLM modelers.
  – Detection of nonlinearity: statisticians devote much energy to addressing potential nonlinearity and threshold effects
  – Outliers and data errors can have large deleterious effects on GLMs but have much less impact on data mining models
  – Statisticians spend much of their time looking for the right set of predictors to use, selecting from a large pool of candidates.
  – Data mining methods can effectively select predictors automatically

• Data mining makes modelers more productive
  – Develop more high quality models in less time
Examples of Data Mining in Action for Health Insurance

- Real world examples that can be publicly reported rare
  - Issues: privacy and proprietary nature of results
  - Can often only report fragments of results released to public
  - Several studies presented at Salford Systems conferences
- Worker’s Compensation: Identifying probable serious cases at time a case is opened
  - WORKCOVER: New South Wales, Australia
  - Analysis conducted by PriceWaterhouseCoopers, Australia
- Lifetime value of a customer
  - Depends on probability of hospital claims and length of stay
- Health related example from automobile injury insurance
Cases Studies
By Users of CART®, MARS®, TreeNet®

- Papers available on request from Salford Systems
  - Charles Pollack B.Ec F.I.A.A. Suncorp Metway, Australia
  - Inna Kolyshkina, Price Waterhouse Coopers, Australia

- Other case studies not included here also available

- CART, MARS, TreeNet, RandomForests® are flagship technologies of Salford Systems
  - Core methods developed by leading researchers at Stanford University and UC Berkeley
  - In use at major banks, insurers, credit card issuers and networks (VISA) and internet portals (Yahoo!)

September, 2008 Copyright © Salford Systems 2008
Case Study: Worker’s Compensation
Predicting Serious Claims at Case Outset

• Minority of claims serious (about 14%):
  – Serious claims are responsible for 90% of costs incurred
  – Case may become chronic (serious) if not managed well early
  – Fast return to work best for insurer and insured
  – Early prediction could accelerate effective medical treatment

• Apply CART to a set of claims to identify variables predicting a serious claim

• 83 variables as potential predictors of “serious claim”

• Categorical predictors with many levels
  – “Occupation code” 285 levels
  – “Injury location code” 85 levels
  – Such variables are handled with ease in CART

September, 2008
Copyright © Salford Systems 2008
• Examples of Data available:
  – About claim:
    • Dates of registration and closing
    • Was the claim reopened?
    • Was the claim litigated?
    • Liability estimates
    • Payments made
    • Was claim reporting delayed?
  – About claimant:
    • Gender, age, family/dependents
    • Employment type, occupation, work duties
    • Wages
  – About injury or disease:
    • Time and place
    • Location on body
    • Cause or mechanism
“Serious Claim” defined as:
- Claimant received payment at least three months (time off work) AND/OR
- Claim was litigated

Modeling based on a random sample of cases
- Injury occurred 18-24 months prior to the latest claim
Results:
- 19 predictive predictors selected from 83 candidates
- Some predictors expected (nature and location of injury)
- Some unexpected (like claimant language skills)

Classified 32% of all claims as serious (test data)

<table>
<thead>
<tr>
<th>Actual/Predicted</th>
<th>Serious</th>
<th>Not Serious</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Serious</td>
<td>6,823</td>
<td>2,275</td>
<td>8,558</td>
</tr>
<tr>
<td>Not Serious</td>
<td>12,923</td>
<td>39,943</td>
<td>52,866</td>
</tr>
</tbody>
</table>
Case Study: Worker’s Compensation
Predicting Serious Claims at Case Outset

• Misclassification tables

**Misclassification for learning data**

<table>
<thead>
<tr>
<th>Class</th>
<th>N Cases</th>
<th>N Misclassified</th>
<th>Percent Error</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Serious</td>
<td>16,922</td>
<td>3,891</td>
<td>22.99</td>
<td>0.23</td>
</tr>
<tr>
<td>Non-Serious</td>
<td>105,358</td>
<td>25,744</td>
<td>24.43</td>
<td>0.24</td>
</tr>
</tbody>
</table>

**Misclassification for test data**

<table>
<thead>
<tr>
<th>Class</th>
<th>N Cases</th>
<th>N Misclassified</th>
<th>Percent Error</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Serious</td>
<td>8,558</td>
<td>2,275</td>
<td>26.58</td>
<td>0.27</td>
</tr>
<tr>
<td>Non-Serious</td>
<td>52,866</td>
<td>12,923</td>
<td>24.44</td>
<td>0.24</td>
</tr>
</tbody>
</table>

– 2/3 data for learning, 1/3 for testing

September, 2008

Copyright © Salford Systems 2008
Case Study: Worker’s Compensation
Predicting Serious Claims at Case Outset

- Model Assessment: Gains chart:

  - Data ordered from nodes with highest proportion of “serious” claims to lowest
  - Baseline is if model gave no useful information
  - Curve is cumulative percentage of “serious” claims versus the cumulative percentage of the total population
  - Difference between baseline and curve is the “gain”
  - The higher above baseline the better the model (larger gain)
Case Study: Modeling Total Projected Customer Value for a Health Insurer

- **Lifetime customer value**
  - Discounted present value of income less associated expenses

- **Develop model for total projected customer value**
  - Multiple sub-models:
    - Hospital claim frequency and cost for next year
    - Ancillary claim frequency and cost for next year
    - Transitions from one product to another
    - Births, deaths, marriages, divorces
    - Lapses
Case Study: Modeling Total Projected Customer Value for a Health Insurer

• Data used for hospital claim frequency and cost sub-model:
  – Covered a 36-month period
  – Predicted outcomes for next 12 months using data from previous 24 months

• About 300 variables as potential predictors:
  – Demographic (age, gender, family status)
  – Geographic and socio-economic (residence location, indices on education, advantage/disadvantage)
  – Membership and product (membership duration, product held)
  – Claim history and medical diagnosis
  – Miscellaneous data (distribution channel, payment method, etc.)
Hospital claim frequency and cost sub-model divided into two sub-models:
  – Predict probability of at least one claim over past 12 months
  – Predict cost given at least one claim

Data segregated with separate models
  – Claims lasting one day
  – Claims lasting more than one day with a surgical procedure
  – Other claims
Exploratory analysis
- Preliminary tree construction to uncover broad groups of data
- CART gave four groups according to age and previous experience

Build separate claims cost models for each group
- Using CART as a model segmentation tool
- Used MARS to build cost regressions

Results
- Similar predictors found among groups (age, hospital coverage type)
- Major differences in models across groups
  • Context dependence
Case Study: Modeling Total Projected Customer Value for a Health Insurer

- Joint CART/MARS 2 stage results
  - The top 15% of members predicted to have highest cost accounted for 56% of total actual cost
  - The top 30% of members predicted to have highest cost accounted for 80% of total actual cost
Case Study: Modeling Total Projected Customer Value for a Health Insurer

- Joint CART/MARS Results: Gains chart

![Gains Chart](chart.png)
Case Study: Modeling Total Projected Customer Value for a Health Insurer

- Two stage model Results:
  Average actual and predicted values for overall annual hospital cost

- Large differential between highest and lowest indicates a good model
- Model follows actual with a good fit
Case Study: Optimizing Premium Increases

- Australia’s 2nd biggest insurer (SunCorp Metway)
  - Modified rates after an acquisition to enforce uniformity
  - Some premiums increased, others decreased (subject to caps)
- Opportunity to study the impact of price changes
- Goal: Identify optimal capping rules for price increases

Graph:
- X-axis: premium change
- Bars indicate frequency among policies
- Blue line is retention rate
- Large premium changes (up or down) lead to lapse

Copyright © Salford Systems 2008
Case Study: Optimizing Premium Increases

- Model 1: Yes/No model for “did customer renew?”
- Data used
  - 12 months of renewal offers. Split 2:1 for training and testing
- Variables included
  - Age of insured
  - Other product holdings
  - Length of time with organisation
  - Distribution channel
  - Geographic Location
  - Age of vehicle/house
  - Method of Payment (Monthly/Annual)
  - Level of ‘No Claims Bonus’
  - Value of vehicle/house
  - Level of Deductible
- Price change not included as it was randomly distributed
Case Study: Optimizing Premium Increases

- Retention tree
- 7 segments
- Excludes price change
Case Study: Optimizing Premium Increases

- Tree translated

Group 1
  - NCD Step Back?
    - Group 2
      - Endorsement?
        - Group 3
          - Risk added mid term? (Renewal term different from last term)
            - Premium Payment Frequency
              - Annual
              - Monthly
            - NCD Level < 40%?
              - Number of previous renewals > 4?
                - NSW, QLD
                - Other
                  - Vehicle Age < 8?
                    - Driver age < 49?
                      - Group 12
                      - Group 13
                    - Group 11
                      - Group 10
            - Group 4
              - Multi-Product Holdings?
                - NCD Level < 40%?
                  - Group 5
                    - Number of previous renewals > 4?
                      - Group 6
                        - CTP Discount?
                          - Group 7
                            - Number of Previous Renewals < 1?
                              - Group 8
                                - Driver age < 42?
                                  - Group 9
                                    - Group 14
                                      - NCD < 40%?
                                        - Group 15
                                          - Business Rules
                                            - Group 8
                                              - Driver age < 42?
Probability of retention as a function of % price change, within CART segment

September, 2008

Copyright © Salford Systems 2008
Probability of retention as a function of $ price change, within CART segment
Case Study: Optimizing Premium Increases

• Results
  – Variable importance differed somewhat from business expectations
  – Notable absence of age of insured from early splits
  – Length of time with company of lower order importance than expected
  – Some variables were important in unexpected ways (like customers with multi-product holdings)
• Does the model work?
  – Even with extremely high cost of new business acquisition, the optimal result is achieved with NO capping
  – Model validated for three months following 12 months data period
    • Predictions matched well with actual results
  – Tree was easily explained to management
  – Some business expectations (myths?) were dispelled
  – Modelling assumptions were validated
Hybrid Case Study: MARS guided GLM

• Data used
  – Industry-wide auto liability data for Queensland, Australia
  – Individual claim data aggregated into the number of claims reported

• Potential predictors include
  – Accident month
  – Number of casualties
  – Number of vehicles in the calendar year
  – Number of vehicles exposed in the month
Hybrid Case Study: MARS guided GLM

• Initial GLM without MARS
  – Poisson model with log link
  – Number vehicles exposed in a month as offset
  – Manual transformation and interactions
  – Assessed with ratio of deviance to the degrees of freedom, predictor significance, link test and residual analysis
  – 5-7 days to generate

• Second GLM based on MARS variables and transforms
  – MARS model
    • ratio of incurred number of claims to number of vehicles exposed in the month as the dependent variable
  – Input resulting MARS basis functions to new GLM (same conditions as initial GLM)
    • Backward elimination to remove a small number of insignificant variables
    • Assessed with same methods as initial GLM
  – One hour to generate MARS-enhanced GLM

• Compare models with assessment results and gains charts
• MARS-enhanced modelling considerably faster and more efficient
• Performance and fit the same

Hybrid Case Study: MARS guided GLM
Hybrid Case Study: MARS guided GLM

- Gains chart
  - Equal performance
  - Gains tables indicate marginally better performance from MARS-enhanced GLM

- High degree of similarity in variable importance

- MARS-enhanced GLM picked up variable interactions not detected by hand-fit GLM
Hybrid Case Study: Retention Modeling

• Data
  – 198,386 records from the UK
  – Each record is one trial / outcome
  – Split 50/50 for training and testing

• 135 potential predictors
  – For GLM each variable is binned
  – 3,752 total levels across all variables

• Combine GLM and CART for one complete model
• Current practice by EMB for casualty insurance GLMs
Hybrid Case Study: Retention Modeling

- GLM (forward regression)
  - 57 significant predictors
  - Took a weekend to run
- CART
  - 24 significant predictors
  - Top 15 shared with GLM
    - Took one hour to run
- Final model has 26 predictors
  - 6 interactions found by CART
  - ROC values of 0.862 (training) and 0.85 (test)
• Combining CART, MARS, and GLM
  – CART: Select predictors, understand data
  – MARS: refine regressors
  – GLM: takes MARS basis functions as predictors

• Can also go from GLM to CART
  – Use CART to analyze GLM residuals

Optimal Model

Refined data set + Important variables

Basis functions

Residuals

Compare with other GLM models

Familiar results format

Familiar statistical analyses

September, 2008

Copyright © Salford Systems 2008
Salford Systems: R&D Staff and Academic Links

- **Dan Steinberg**, PhD Econometrics, Harvard (Data Mining)
- **Nicholas Scott Cardell**, PhD Econometrics, Harvard (Data Mining, Discrete Choice)
- **Jerome H. Friedman**, Stanford University (algorithm coder CART, MARS, Treenet, HotSpotDetector)
- **Leo Breiman**, UC Berkeley (algorithm developer, ensembles of trees, randomization techniques to improve trees)
- **Richard Olshen**, Stanford University (Survival CART, Tree-BasedClustering)
- **Charles Stone**, UC Berkeley (CART large sample theory)
- **Richard Carson**, UC San Diego (Visualization Methods, Super Computer methods)
Salford Systems: Selected Awards

- 2007 Winner of the DMA Analytics Challenge (targeted marketing)
- 2007 Grand Champion for the PAKDD Data Mining Competition
- 2006 First runner-up for the PAKDD Data Mining Competition
- 2004 First place for the KDD Cup (accuracy in particle physics)
- 2002 Winner of the Duke University/NCR Teradata CRM center data mining and modeling competition
- 2002 Jerome Friedman (developer of CART, MARS, TreeNet) awarded the ACM SIGKDD Innovation Award
- 2000 Winner of the KDD Cup 2000 International Data Mining competition
- 1999 Deming Committee winner of the Nikkei Prize for excellence in contributions to quality control in Japan
Salford Systems: Contact information

• Contact us to obtain the studies on which these slides were based

• Salford Systems world headquarters
  – info@salford-systems.com
  – 4740 Murphy Canyon Rd. Suite 200
  – San Diego CA, 92123
  – (619) 543-8880 (voice)
  – (619) 543-8888 (FAX)