Data Preparation for Predictive Modeling

2008 CAS Ratemaking Seminar
March 17-18, 2008

John Stenmark
Consulting Actuary
Actuarial Data Management Services

Is this You?

Data Preparation for Predictive Modeling

One-way analyses vs. Multivariate Analyses

- Traditionally, data preparation for actuarial analysis involved the aggregation of data into summaries.
  - Relatively small amount of data (number of data dimensions were additive)
  - Data anomalies were buried in the aggregation.
  - Rarely impacted the analysis significantly…
  - …at least that's what we thought.

Data Preparation for Predictive Modeling

One-way analyses vs. Multivariate Analyses

- Predictive Modeling requires data at a far more granular level.
  - Greater amount of Data (number of dimensions are multiplicative)
  - Data anomalies are more obvious.
  - Data anomalies may hinder or even prevent the analysis.
  - Data is still aggregated at some level and great consideration must be given in the determination of banding definitions.
  - External data sources.

Data Preparation for Predictive Modeling
Ravi Kumar

- Ravi Kumar is an Associate of the Casualty Actuarial Society and a Member of the American Academy of Actuaries.
- He is a Senior Manager in Deloitte Consulting's Actuarial and Insurance Solutions practice, where he has built numerous predictive models for personal and commercial lines.
- Prior to joining Deloitte, he worked at NCCI, Boca Raton where he designed and developed large actuarial applications for Ratemaking, Reserving and Benchmarking.
- Ravi holds Bachelors in Engineering from IIT, Madras India and a Masters in Computer Engineering from Florida Atlantic University, Boca Raton.

Data Preparation for Predictive Modeling
Wayne Jiang

- Wayne Jiang is a Fellow of Casualty Actuarial Society.
- He joined SAFECO Insurance Companies in 1999 where he worked on reserving from 1999 to 2001. He has been working on building predictive models for commercial lines insurance since 2001. He is currently a director in SAFECO business insurance.
- Prior to joining SAFECO, he worked for AON Corporation for pricing and reserving.
- He obtained B.S. from University of Science and Technology of China in China and M.S. from University of Washington.

Data Preparation for Predictive Modeling
John Stenmark

- Fellow of the Casualty Actuarial Society and a Certified Insurance Data Manager
- Over thirty years with Southern Farm Bureau Casualty Insurance Company (Senior Vice President – Actuary).
- Chairman of the Casualty Actuarial Society's Committee on Management Data and Information.
- Ratemaking Seminar Committee and is responsible for the Data Track of the 2008 Ratemaking Seminar.
- Several presentations to the CAS on various Topics including a paper Co-authored by Peter Wu: Simpson's Paradox, "Confounding Variables and Insurance Ratemaking," and panels entitled "Data Warehousing on the Cheap," "Information Stored, Mined and Utilized" and "Data Preparation for Predictive Modeling."
- Recently formed Actuarial Data Management Services, a consulting organization that helps companies fulfill their actuarial data requirements.

Data Preparation for Predictive Modeling

Overview

- Discussion of the Data Requirement differences between one-way analyses and Multivariate Analyses
- Types of Data Variables
- Discussion of Data Quality Issues
- Some Parting Tidbits

Data Preparation for Predictive Modeling

Just to be clear!!

- Univariate Stats (for a territory analysis) might look like this:

Year	Terr	Premium	Exposure	Losses	Claim Count
2006	01	423,295	756	321,754	24
2006	02	216,500	500	151,550	16
2006	03	618,768	1,206	525,876	38
2006	04	159,750	355	116,618	11
2006	06	496,850	950	248,425	30

- But Multivariate Analysis requires stats that might look like:

Year	Pol. #	ZIP	Terr	Class	Premium
2006	1234	39110	22	1A	120.00
2006	3975	39110	22	1A	60.00
2006	4245	39158	23	1A	70.00
2006	9573	39158	23	1A	35.00
2006	4519	39158	23	1B	140.00

Data Preparation for Predictive Modeling

- So for Predictive Modeling we're talking about significantly more data than with One-way Analyses and …
- … the Quality of the data is far more Critical.
- A data cleansing stage is required between processing systems and Predictive Modeling software.
- A Database (or even better a Data Warehouse) is necessary to conduct the pre-analysis to ensure crisp data.
- The cleansing should be done so as to benefit other Projects such as Competitor Monitoring, Rate Impact Analysis, Loss Reserve Analysis and Catastrophe Modeling.

Data Preparation for Predictive Modeling

Types of Data Variables

- Critical Variables
  - Data that without which the process fails
- Current Rating Variables
  - If you can't price your current structure you can't sell your Predictive Modeling Project
- Other Internal Variables
  - Aids the Modeling process
  - Provides future Rating Variables
- External Data
  - This puts the "Predictive" into Predictive Modeling
- Internal External Data
  - For example, agent information, Auto experience for predicting Homeowners, etc.

Data Preparation for Predictive Modeling
Critical Variables

- The Facts – Exposures, Premium, Claim Counts, Losses (Normally Premium is the least critical for Predictive Modeling)
- Unique Policy/Claim, Unit (vehicle, location) and Coverage Identifiers
- Essential Dates
  - Policy Effective Date
  - Coverage Effective Date
  - Transaction Effective Date

Data Preparation for Predictive Modeling
Rating and other Variables

- Critical to the acceptance of Predictive Modeling Process
- Include underlying variables
- I. e. Include Age, gender, marital status as separate variables as well as class.

Data Preparation for Predictive Modeling

Data Quality Issues

- Data Quality – whose responsibility is it?
- Whose Job Performance Suffers because of Bad Data?
- Whose Life is Miserable because of Bad Data?
- Yours!!!

Data Preparation for Predictive Modeling

Data Quality Issues

- Data quality is critical.
- More than just balancing and edits.
- Graphical Representation is crucial
  - Identify miscoding
  - Identify missing values
- Since many variables are involved record keeping (logging) is crucial
- The source for Predictive Modeling Data should be at the transaction level

Data Preparation for Predictive Modeling
Data Quality Issues
The Five Misses

- Misclassification
- Miscoding
- Misinterpretation
- Missing values
- Miscalculation

Data Preparation for Predictive Modeling

Data Quality Issues

- The most serious problems with Predictive Modeling Data are caused by mismatches in Numerator and Denominator.
- Assume a very granular data base (only one policy per cell).
- Consider if claims data is coded to zip code 39110 and the policy is coded to 39158. The frequency for the 39158 cell will be zero and the frequency for the 39110 cell will be undefined (divide by zero).
- Infinite frequencies somewhat confuse most predictive modeling systems.

Data Preparation for Predictive Modeling

Data Quality Issues

- The same thing can happen to severity.
- If each time there is a claim payment the policy master file is queried and only rating variables are time stamped then a change in zip code would cause a similar disconnect between claim payments and claim count.
- In fact if there are reopen claims there might be negative claim counts generated for some claims.

Data Preparation for Predictive Modeling

Data Quality Issues

- Mismatches in numerator and denominator can happen when stats are not properly generated for midterm changes when some policy characteristic that is not also a rating variable is involved.
- When a rating variable (e.g. Class) changes midterm, generally that also changes the premium. The onset and offset transactions keep the reported stats for that variable correct.
- For example, assume the insured moves midterm from zip code 39110 to 39158. Say this moves the insured from territory 22 to 23.
- Assume that later in the policy period the class changes from 1A (adult) to 2A (youthful male).

Data Preparation for Predictive Modeling

Data Quality Issues

- The stats including the original premium might look like this:

Trans Dte.	Pol. #	ZIP	Terr	Class	Premium	Annual PIF	Exposure
2007-01-15	1234	39110	22	1A	120.00	120.00	+1.00
2007-07-15	1234	39110	22	1A	-60.00	120.00	-0.50
2007-07-15	1234	39158	23	1A	+70.00	140.00	+0.50
2007-10-15	1234	39158	23	1A	-35.00	140.00	-0.25
2007-10-15	1234	39158	23	1B	+140.00	560.00	+0.25

- But if the change in Zip Code doesn't alter the territory (say if ZIP 39110 and 39158 are both in territory 22) some processing systems might not generate any new stat records on 7/15.

Data Preparation for Predictive Modeling

Data Quality Issues

- The stats including the original premium might look like this:

Trans Dte.	Pol. #	ZIP	Terr	Class	Premium	Annual PIF	Exposure
2007-01-15	1234	39110	22	1A	120.00	120.00	+1.00
2007-10-15	1234	39158	22	1A	-35.00	140.00	-0.25
2007-07-15	1234	39158	22	1B	+140.00	560.00	+0.25

- Notice that while the exposures are correct for 1A all the premium and exposures are booked only to Zip Code 39110. This is incorrect, but not tragic, but what if there is a claim on 11/1? If the claims system uses the updated master file to post the ZIP Code as 39158 then a frequency for 39158 will be created with an undefined value.

Data Preparation for Predictive Modeling

Data Quality Issues

The Long Term Solution

- Each Data element change must generate a stat record with an onset and offset premium and Annualized Premium In Force.
- The stats including the original premium would look like this:

Trans Dte.	Pol. #	ZIP	Terr	Class	Premium	Annual PIF	Exposure
2007-01-15	1234	39110	22	1A	120.00	120.00	+1.00
2007-07-15	1234	39110	22	1A	-60.00	120.00	-0.50
2007-07-15	1234	39158	22	1A	+60.00	120.00	+0.50
2007-10-15	1234	39158	22	1A	-30.00	120.00	-0.25
2007-10-15	1234	39158	22	1B	+140.00	560.00	+0.25

Data Preparation for Predictive Modeling

Data Quality Issues
The Short Term Solution

- Data must be cleansed.
- A Policy (or Claim) / Unit (Vehicle, Location, etc.) profile must be derived for each transaction period and used throughout the Transaction effective period.
- The stats including the original premium must look like this:

Trans Dte.	Pol. #	ZIP	Terr	Class	Premium	Annual PIF	Exposure
2007-01-15	1234	39110	22	1A	120.00	120.00	+1.00
2007-10-15	1234	39110	22	1A	-35.00	140.00	-0.25
2007-07-15	1234	39158	22	1B	+140.00	560.00	+0.25

Data Preparation for Predictive Modeling

Data Quality Issues

One Comment on Data Quality

- The statement has been made; "As long as the agents keep misclassifying business it is futile to attempt to collect valid stats."
- Valid stats that reflect precisely the business as it is written is the best weapon against future misclassification.
- Valid stats that reflect precisely the business as it is written is the best source of ratemaking data. I.e. rates derived will reflect that a past misclassification will be repeated in the future.
- This is true even to the extent that if you could cleanse this data you shouldn't unless you anticipate that the agents will cease misclassification in the future.

Data Preparation for Predictive Modeling

External Data Sources

- Predictive Modeling allows/encourage the use of data outside of the rating variables and, in fact, outside of the company.
- The first external data that a company is likely to use is Demographic Data, usually by Zip Code.
- The source for most of these data is the US Census Bureau at http://www.census.gov/ . Some possible variables appear below:

- Average Education Years
- Population Density
- Mean Age
- Percent Rural
- Percent Farm
- Travel Time
- Median Income
- Median year Owner occupied structure built
- Median year householder moved into unit
- Median value for all owner-occupied housing units
- Median price asked
- Median selected monthly owner costs

- To find out about these attend the session "External Data Sources."

Data Preparation for Predictive Modeling

Some Parting Tidbits

- The key to Quality Data is a Data Quality philosophy that goes to the top level of management. To find out how to do that, I recommend hiring a full time data manager that has passed or will pass the Insurance Data Management Association (IDMA) exams. The exams cover: 1. Insurance Data Collection and Reporting, 2. Insurance Data Quality, 3. Systems Development and Project Management and 4. Data Management, Administration and Warehousing.
- Attend the session: Raising Your Actuarial IQ (Improving Information Quality)

Data Preparation for Predictive Modeling

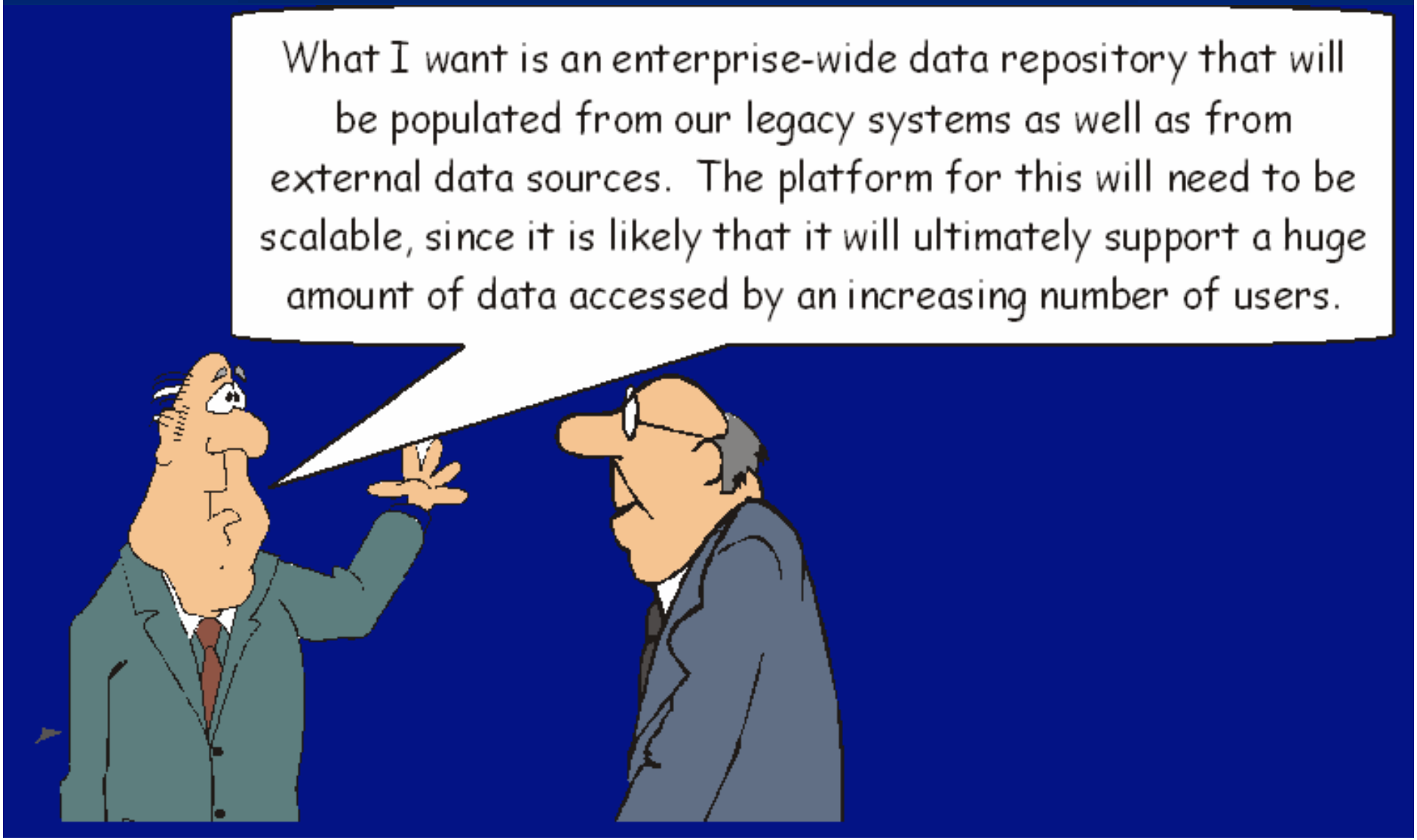
Some Parting Tidbits

- Speaking of Data Quality how about Agent's incentives that promote accuracy in data. Production (especially on the Gulf Coast) is out, loss ratio is important, but in the twenty-first century Quality Data is more important. How about agent's trips based on Data Accuracy.
- Be wary of companies that promise an "Actuarial" Data Warehouse.
  - Most of these companies produce data marts or cubes. These are perfect for traditional one-way analyses, but you are dead in the water if you try to use them for driving your predictive modeling data requirements.
  - They are not appropriate for producing data for Cat Modeling or batch processing Competitive Monitoring Reports (Quadrant)

Data Preparation for Predictive Modeling

Communications

The Actuary to the Data Warehouse Vendor



Data Preparation for Predictive Modeling

2008 CAS Ratemaking Seminar
March 17-18, 2008

John Stenmark
Consulting Actuary
Actuarial Data Management Services
(601) 955-3022
jstenmark@comcast.net