Text Mining on Unstructured Data

Presented at
CAS Ratemaking and Product Management Seminar
March 11, 2009 (Las Vegas)

Presented by
Philip S. Borba, Ph.D.
Milliman, Inc.
New York, NY

PRESENTATION OBJECTIVES AND OVERVIEW

- Starting Principles for Predictive Models
- Overview of Text Data for Predictive Models
- Use of Text Data in Predictive Models

STARTING PRINCIPLES FOR PREDICTIVE MODELS

- 2 Major Classes of Models
  - Macro models
  - Micro models

- General Structure of a Predictive Model
  - Outcome = f(data)

- 3 Key Elements in Predictive Models
  - Outcome measures
  - Data
  - Analytical techniques

STARTING PRINCIPLES:
2 Major Classes of Models

- Macro Models
  - Outcome measure = result for a group of policies or claims (industry, company, division, book of business)
  - Analyses to evaluate, detect, or seek trends or patterns

- Micro Models
  - Individual account or claim results
  - Can be rolled up to a macro analysis but need micro data for the starting point

- Present focus: Micro Models

STARTING PRINCIPLES:
3 Key Elements in Predictive Modeling

- Outcome measures
  - What result am I measuring?
  - Loss ratio, loss cost, claim frequency, severity

- Data
  - What data are available to analyze the outcome?
  - Considerable amount of new data

- Analytical techniques
  - What is an appropriate analytical technique? (may be more than one)
  - Cost of additional analytical sophistication may not warrant additional complexity

STARTING PRINCIPLES:
3 Key Elements in Predictive Modeling

- Outcome measures
  - Over the past 25 years(at least), not a significant change in the types and nature of outcome measures

- Analytical techniques
  - Increased sophistication over past 25 years but generally same set of tools

- Data
  - Element with the most change
  - Principal focus for the balance of the presentation

PRESENTATION OBJECTIVES AND OVERVIEW

- Starting Principles for Predictive Models
- **Overview of Text Data for Predictive Models**
- Use of Text Data in Predictive Models

DATA

- 3 general classes
  - Structured
  - Semi-Structured
  - Unstructured
- Text Data
  - Semi-Structured
  - Structured
- Semi-Structured
- Unstructured

DATA: THREE GENERAL CLASSES

- Structured
  - Examples: claim master record, payment transaction records
  - Fixed format
  - Assigned values (eg, “M” indicates married)
  - Stored in data warehouses, system of data tables or files
  - Able to link files using policy number, claim number, or some other identifier
- Semi-Structured
  - Claim descriptions, payee name and description on transaction records
- Unstructured
  - Examples: claim adjuster notes, police reports
  - Free format, length not fixed
  - Little consistency in raw data (eg, IV RE OV can be expressed may ways)

DATA: STRUCTURED

- Structured
  - Demographic characteristics (eg, age, sex, occupation, marital status)
  - Claim status, class identification
  - Dates of loss, claim reporting, status changes
  - Transactions (eg, payments, reserves)
  - Situation flags (eg, attorney representation, controversion)
  - External data
    - US Census
    - Dun & Bradstreet
    - Commercial compilations of publicly available data
  - Use zip code, SIC, or other characteristics for information on location, economic conditions, economic profile (uses for WC, auto, HO, mortgage)

TEXT DATA

- Definitions:
  - “Text File” – a computer file consisting solely of printable characters from a recognized character set
  - “Plain Text” -- an ordinary "unformatted" sequential file readable as textual material without much processing
- Two styles
  - Semi-structured
  - Unstructured

TEXT DATA

- Free form provides quicker access to information than structured data
  - Suspicious claim (identified before a SIU payment)
  - Recovery opportunities
  - Attorney involvement (before payment to attorney; not relying on 0/1 switch)
  - Claim severity (before payments reach a particular threshold)

- Capturing new information in structured data may require:
  - System changes to accommodate new values for existing fields (eg, new causes of injury or occupational diseases for WC)
  - System changes for new fields

TEXT DATA

- Ability to gather data that may be difficult to capture in a structured data format:
  - Was driver using a GPS?
  - Was driver distracted by a GPS?
  - Was driver using a cell phone?
  - Was driver distracted by a cellphone?
  - Has a payment been authorized for a service that has not been performed? (eg, IMEs that have not been completed)
  - Why are claims remaining open so long?
  - Was alcohol a contributing factor in the accident?

TEXT DATA

- In common use in other businesses
  - Internet search engines
  - Security applications
  - Online media relationships
  - Marketing applications
  - Indexing
  - Spam filters
  - Categorization of movie reviews
    - Positive: “dazzling”, “brilliant”, “excellent”, et. al.
    - Negative: “terrible”, “awful”, “hideous”, et. al.

TEXT DATA: SEMI-STRUCTURED

- Characteristics
  - Text with a limited record length or entered in a “description” field

- Types of files
  - Case narrative / claim description (often 30-100 bytes, free form)
  - Payment transactions (payee name, payee address, payment description)
  - Police reports
  - Portions of adjuster notes
  - Portions of nurse case management notes
  - Auto, home, et. al. appraisals

TEXT DATA: UNSTRUCTURED

- Types of files
  - Adjuster notes
  - Nurse case management notes
  - Appraisal reports
  - Auto-, home-, et. al. repair reports
  - Depositions, court transcripts
  - e-mails
  - Memos and letters
  - pdf files; flat-text files
  - Underwriter notes
  - Policy notes
  - Safety-inspection records

TEXT DATA: UNSTRUCTURED

- Challenges in accessing
  - A single “record” can be 100s, 10,000s, or more bytes
  - pdf files; flat-text files
  - HTML format
  - Unnecessary “bloat” information
    - Certain phrases may be stock entries that provide no information
    - Formatting keys that provide no information
  - Single file with several types of unstructured data
    - Policy underwriter, first report of injury, claim adjuster notes, and claim appraisal information in a single file

TEXT DATA: UNSTRUCTURED

- Three uses:
  - A single textual reference initiates an action
    - Claims are queued for a claim administrative action
  - A pattern in the text information initiates an action
    - Subrogation, SIU, IME
  - Creation of additional fields for modeling
    - Probability of an event (claim frequency) (eg, suspicious claim, perm partial disability)
    - Cost of an event (severity) (eg, total losses)

TEXT DATA: UNSTRUCTURED

- Challenges in processing
  - Extracting useful information / judiciously excluding unneeded information
  - Establishing rules for causation (eg, OV re IV, OV re by IV)
  - Building dictionary of synonyms
  - Building a system to extract desired information
    - Casual direction
  - Building a system to “score” claims
    - How good is the recovery opportunity?
    - What is the liability apportionment?
  - Determining a priority for conflicting information
    - Latest information not always the desired or most suitable for scoring
  - **DO NOT TRY THIS WITHOUT A SME (SUBJECT MATTER EXPERT)**

TEXT DATA: UNSTRUCTURED

- Situations for using unstructured data
  - Casual (eg, IV re OV, OV re IV)
  - Working conditions for WC injuries
  - Lengthy description of accident and circumstances
  - Responsible party for auto accidents
  - Apportionment of liability
  - Circumstances for certain types of claims
    - Multiple claimants for a single claim
    - Multiple injuries for a WC claim
    - Second-injury opportunities for a WC claim
    - Water damage (esp., hurricanes, floods, broken pipes)

PRESENTATION OBJECTIVES AND OVERVIEW

- Starting Principles for Predictive Models
- Overview of Text Data for Predictive Models
- **Use of Text Data in Predictive Models**
  - **Analytical analyses**
  - **Multivariate analyses**

TEXT DATA IN PREDICTIVE MODELS

- Univariate analyses
  - Analyses that look to group phrases that perform as synonyms
    - Example: OV re IV, #2 re #1 (do not include “IV re OV”)
  - Data classification (eg, Pharmacy transactions)
    - Pharmacy transactions
    - Suspicious claims
  - ANOVA
    - Summary statistics for phrases within a population
    - Summary statistics for comparisons across populations
  - Scoring techniques
  - Data-segmentation analyses
  - Cluster analyses

TEXT DATA IN PREDICTIVE MODELS

- Multivariate analyses
  - Outcome = f (structured data, text data)
  - 0/1 variables are created for phrases that perform as synonyms
  - 0/1 variables created in Factor, Principal Components, Cluster, et. al. analyses
  - 0/1 variables included in multivariate models
  - Outcome = f(structured data, GPS usage, cell phone usage)
    - where Outcome = claim cost
    - where Outcome could be 0/1 for rear-end accident

PRESENTATION SUMMARY

- Starting Principles for Predictive Models
  - Outcome = f(data)
  - Data
  - Analytical Techniques

- Overview of Text Data for Predictive Models
  - Structured
  - Semi-Structured
  - Unstructured

- Use of Text Data in Predictive Models
  - Univariate techniques
  - Multivariate techniques