Text Mining for Insurance claim cost prediction

Inna Kolyshkina
PricewaterhouseCoopers

Agenda

- Introduction: why Text Mining now?
- A walk through the Text Mining process
- Quantifying the benefits realised
- Challenges overcome and lessons learned
- Summary

Emerging need for text mining

- Up to 80% of data stored by organisations is in the free text form.
- The data that is contained in text fields holds huge untapped value, BUT free text fields cannot be included in a model!

TEXT MINING is a process that solves this issue:

- “translates” text into numeric form by extracting the patterns from natural language text
- allows us to directly incorporate textual information into predictive modelling.

Some industry applications

- Claims cost prediction
- Fraud and overservicing detection and prevention
- Identification of emerging issues and development of relevant prevention strategies
- Use call centres logs for churn prediction and customer satisfaction measurement

Case Study. Client: Major Australian Insurer

Client wanted:

1. To determine whether the use of unstructured text data in the claim documentation could be used:

(a) to improve existing models of claim cost prediction

(b) to enhance the existing injury coding system.

2. Assistance in making decision regarding investment in text mining software

Our approach

Case Study Context

Task: predict, using data available at the time of the incident report, whether the incident would result in a weekly top 10% claim pay-out value by the end of the next quarter.

Data used: open claims with 18 month history.

Stage 1. Using client data, assess whether textual information has predictive value in predicting claim cost

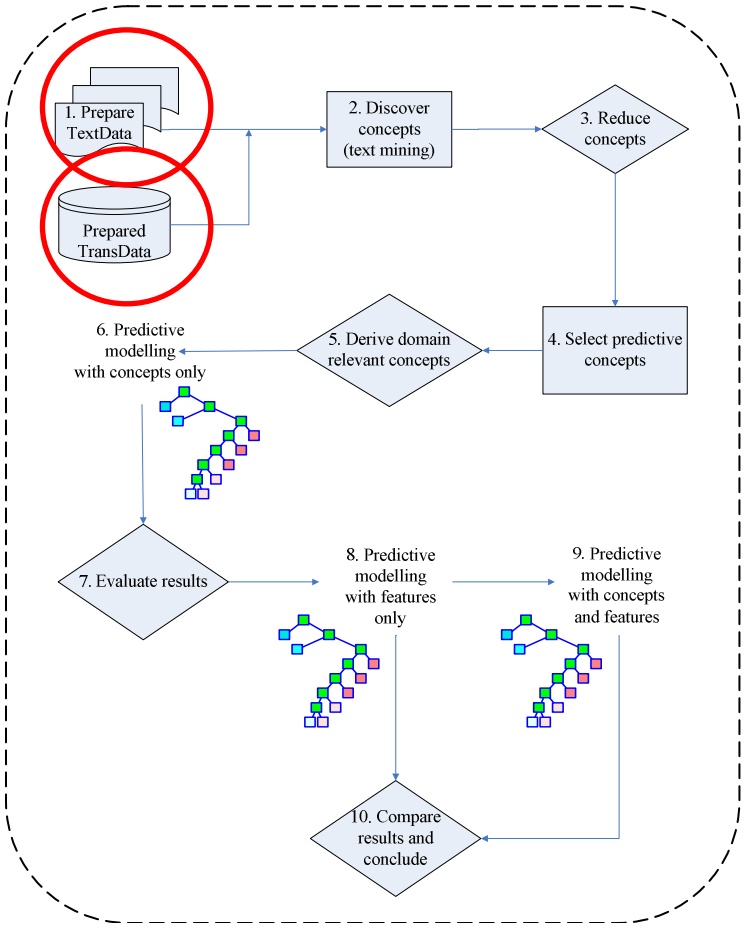
Stage 2. Given that the answer at stage 1 is “yes”, assess whether textual information adds value to the existing models

Text Mining Process.

Step 1	Prepare data
Step 2	Text Mining. Discover concepts
Step 3	Reduce concepts
Step 4	Build data with textual data. Select predictive concepts
Step 5 - 7	Derive domain-relevant concepts. Predictive modelling with text (concepts) only. Assessing whether text is predictive of claim cost
Step 8-10	Assess: does incorporation of text data add value to existing models?

Text Mining Process.

Step 1

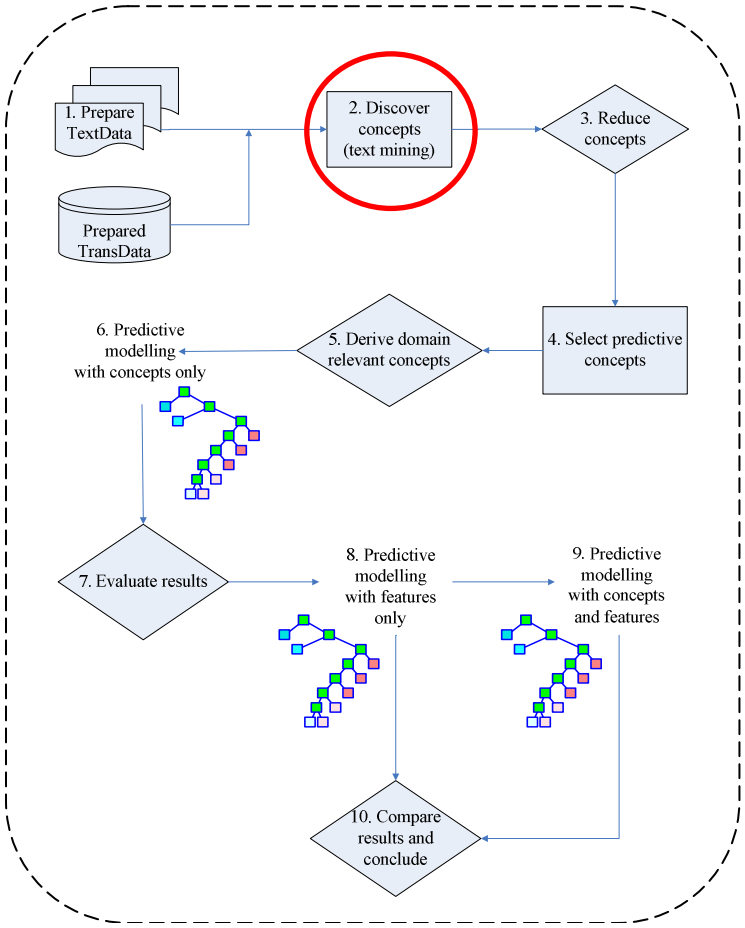


Data Source and Preparation

- Open claims with 18 month history. Approx. 56,000 records.
- **Traditional data:** demographics, payment information and incident codes.
- **Text Data:** Unstructured text fields (~200 chr) about the incident and resulting injury.
- **Target variable for prediction** - Binary indicator if the injury report had a claim pay-out value within the top 10 percent by the end of the next quarter.

Text Mining Process.

Step 2

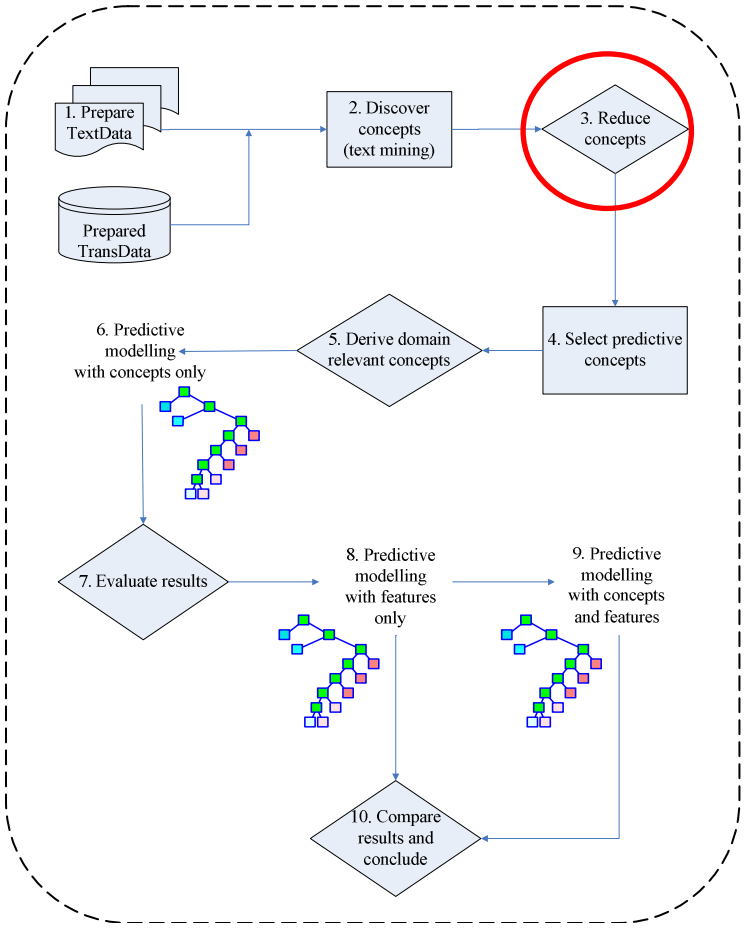


Text Mining. Discover concepts

- A **Concept** is a word or combination of words resident in the text.
- The process required domain business expertise and software package knowledge.
- The process was iterative experimentation to find the optimal algorithm settings.
- Algorithm settings included language and mathematical weightings.

Text Mining Process.

Step 3

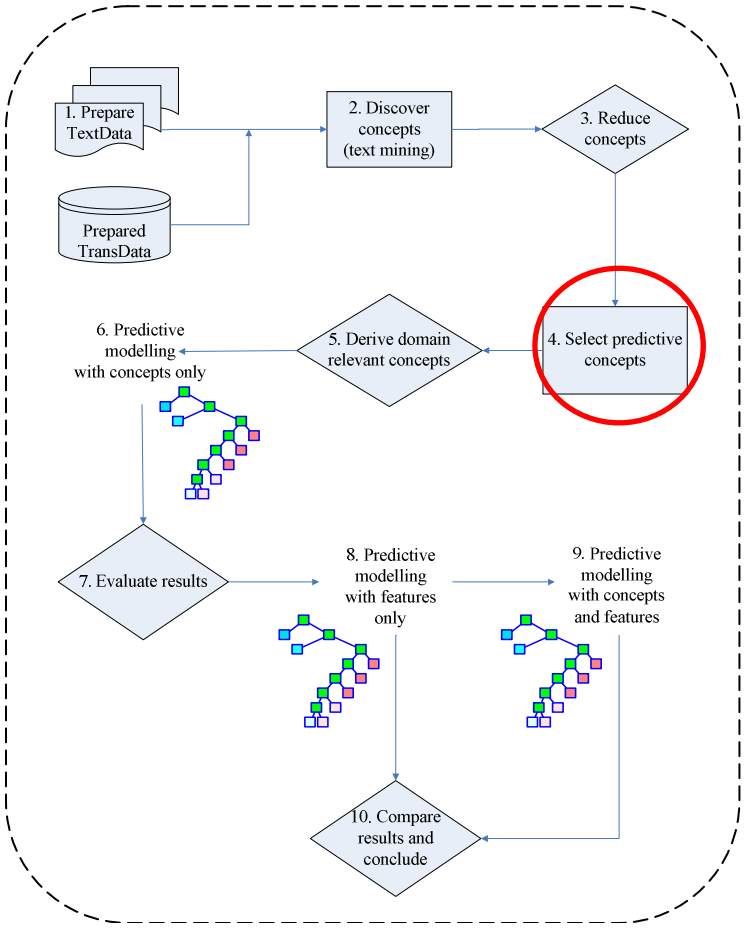


Reduce the number of concepts.

- 8000 concepts were discovered.
- Difficult to make sense of so many concepts
- Concepts with a low frequency would not be relevant within our context.
- Researchers filtered out those concepts which had a frequency of <50.
- After filtering 860 concepts remained.

Text Mining Process.

Step 4



Build data with textual data. Select predictive concepts

Use TreeNet® to identify the 60 most predictive concepts of the remaining 860 concepts.

Concept	Importance
Leg	**100**
Lacerated	**99.43**
Fracture	**92.56**
Stress	**92.27**
Eye	**86.56**
Hernia	**84.11**
Truck	**82.62**
Burn	**73.06**
Ladder	**58**

TreeNet® Overview

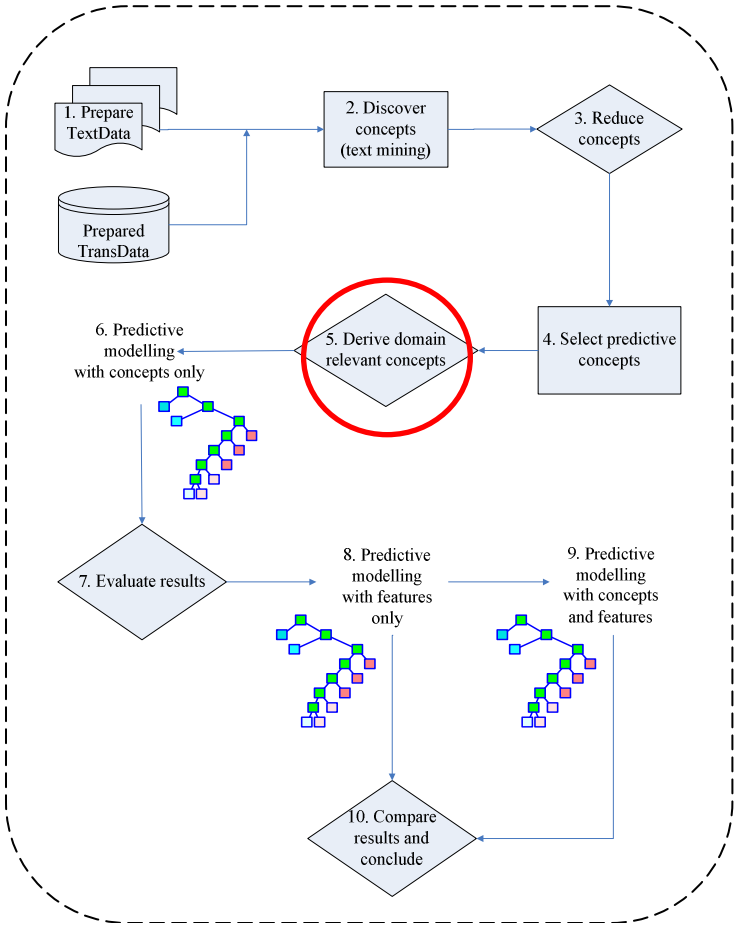
- Model normally consists of several to several hundred smaller summed and weighted trees.
- Trees typically smaller than two to eight terminal nodes.
- Similar to long series expansion (Fourier/Taylor's series)
- A sum of factors that becomes progressively more accurate as the expansion continues – as shown by the equation:

$$F(X) = F_0 + \beta_1 T_1(X) + \beta_2 T_2(X) + \ldots + \beta_M T_M(X)$$

- T_x is a small tree.
- The first tree contributes the most to the model, while subsequent trees contribute successively smaller corrections.

Text Mining Process.

Step 5

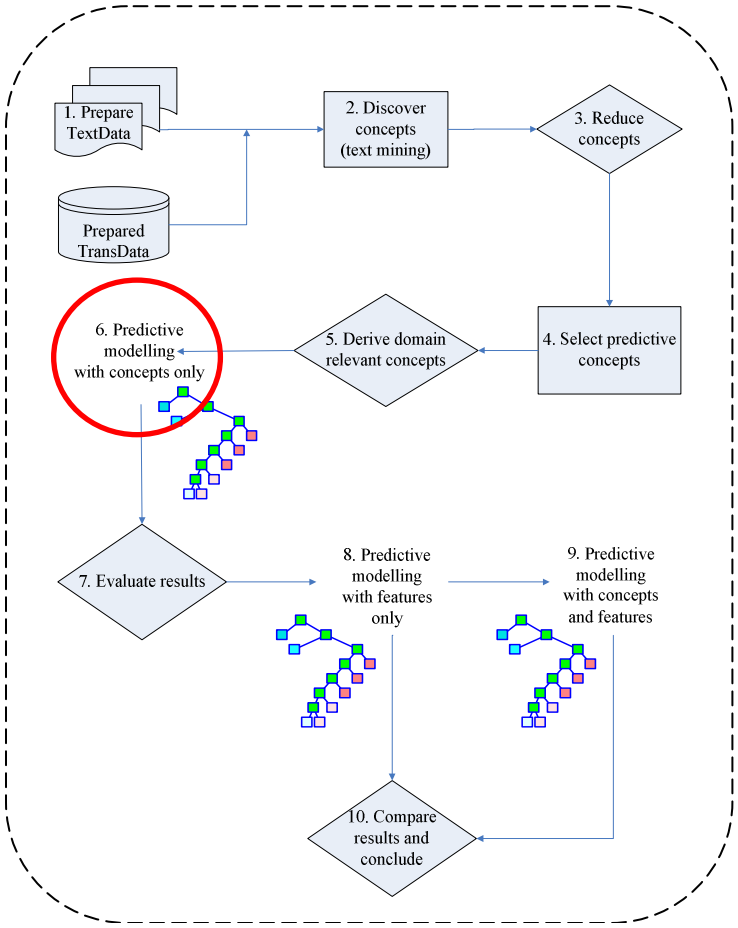


c

- Incorporation of insurance domain expertise.
- Add domain expertise-derived features.
- Grouped concepts with those with a similar meaning, eg stress = anxiety, laceration = abrasion

Text Mining Process.

Step 6



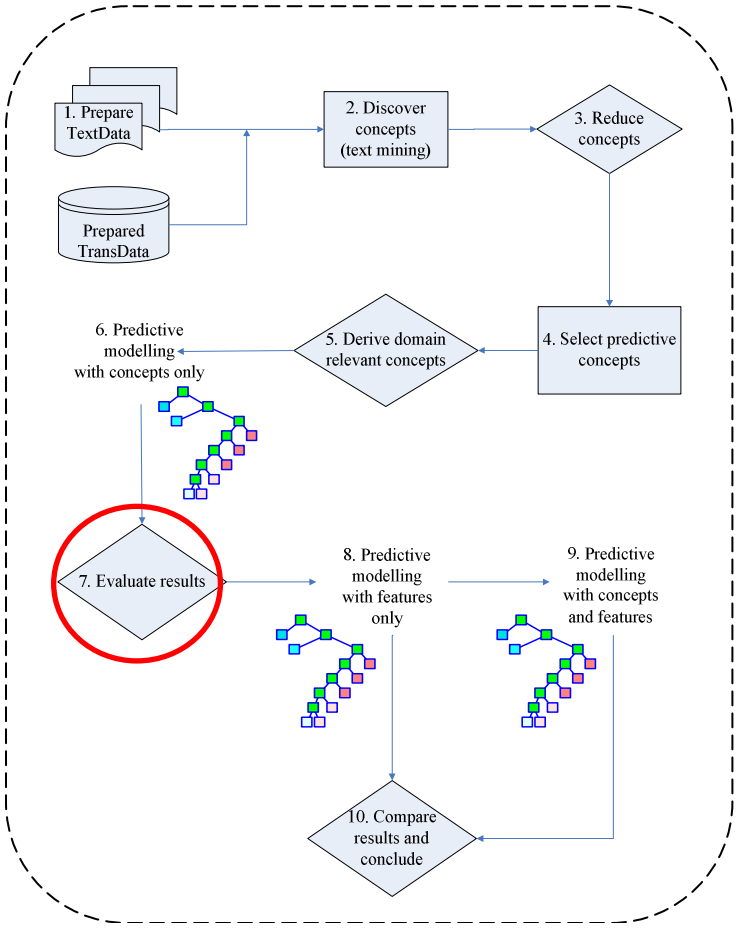
Predictive modelling with text (concepts) only

Built **CART®** predictive model for claims cost using as predictors:

- the concepts identified by **TreeNet®**
- the derived concepts

Text Mining Process.

Step 7

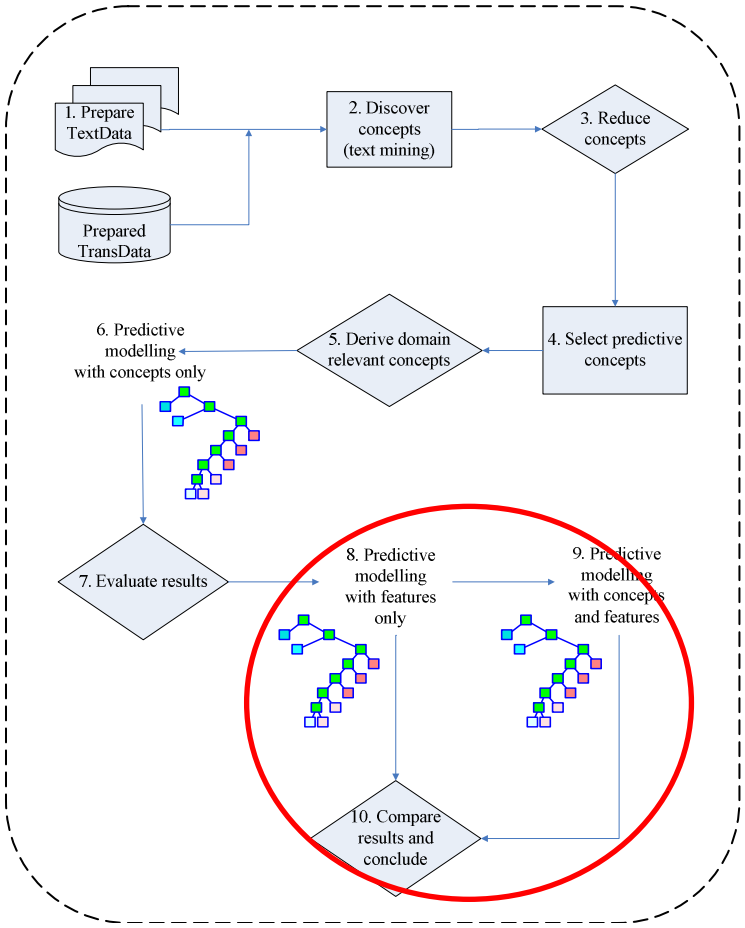


Evaluate model results

- Evaluated models based on **Gains charts** and **model precision measures**.
- The TreeNet® model using concepts only was **75.7%** accurate on test data.

Benefits - Measurement

Steps 8-10

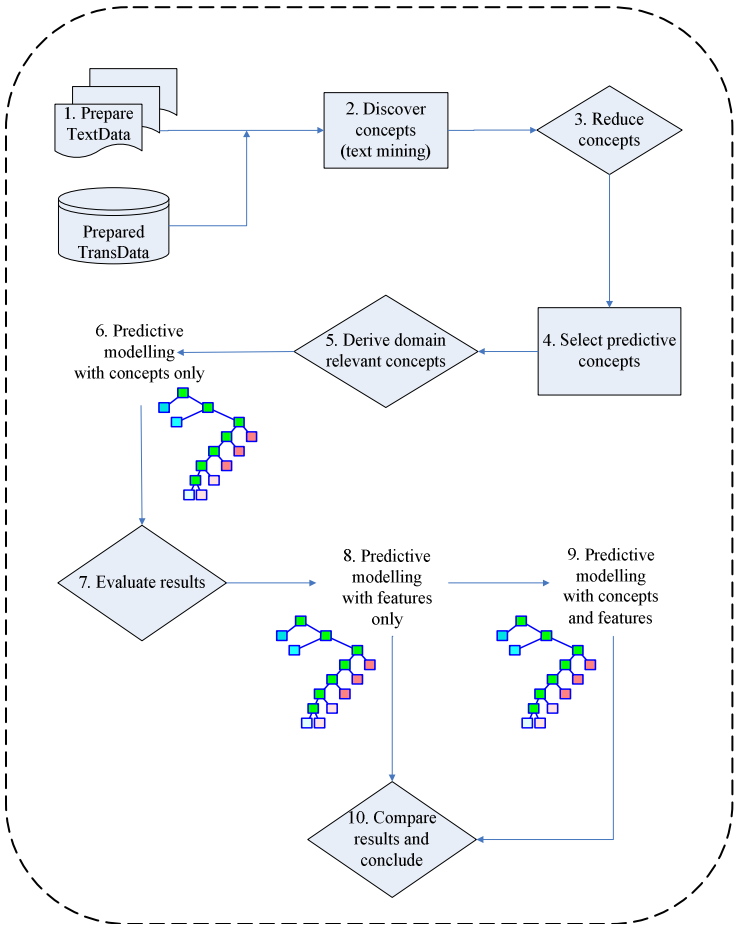


Assessing: does text add value to existing model?

- Created models with demographic and injury codings information only
- Compared them to the models with added textual information.

Text Mining Process.

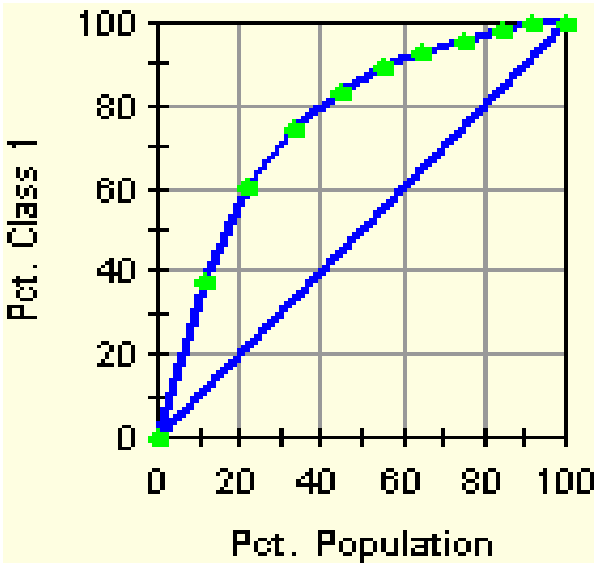
Recap



- 8000 concepts discovered. Removed concepts with a frequency of >50
- 860 concepts remained. Used **TreeNet®** decision trees to select most important concepts.
- Used domain expertise to enrich concepts
- Built predictive model using textual information only. Model was correct in 75.7% cases!
- Adding textual information improved existing models

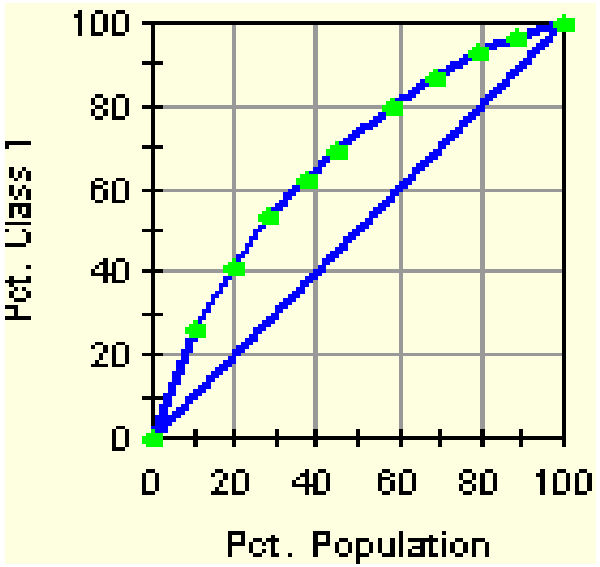
Measurement – Gains chart

- Along the horizontal axis of the gains chart is the percentage of the claims ranked as the most likely to become expensive by the model. Along the vertical axis is the percentage of actual expensive claims appearing in group corresponding to the value of the horizontal axis.
- The ability of the decision tree to segment the data can be measured by the means of a gains chart.
- The distance of the curve above the line y=x gives a measure of the model. An ideal gains chart would rise very quickly to 100%. A poor gains chart would remain very close to the y=x line.

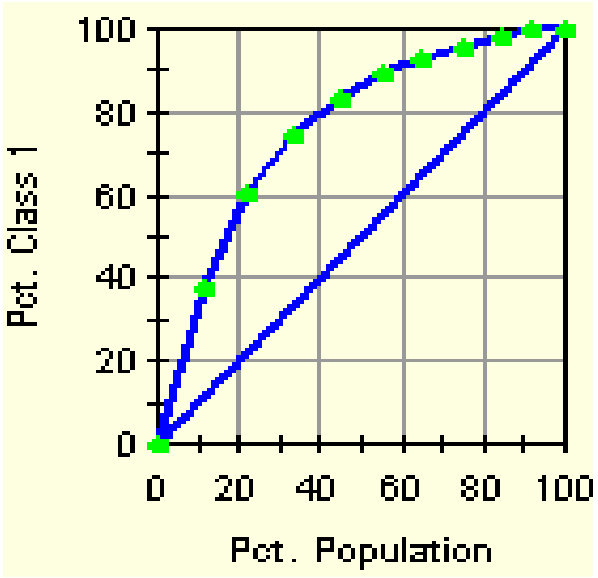


Benefits - Text Mining provides greater predictive power

Claim cost for sprains for the next 6 months (top 5%)



Without



With

Benefits – Identification of high significance concepts not covered by codes

Using the textual data we identified additional highly predictive concepts, which were not existing systemic injury code options. Examples were:

- Box
- Truck
- Injury
- Ground

Benefits – Business Benefits

- Showed that client's existing accident coding system can be enhanced using free text
- Increased precision of claim cost prediction
- Identified the capability of text mining and how it could be used for improvement in other areas of the business
- Assist to decide on investing in a commercial text mining software package

Challenges and Lessons Learned

Data quality

- Computer-generated text integration with human-created text
- Spelling issues (“received” vs “recieved”)
- Abbreviations (“rec” stands for “receipt” and “record”)
- Synonyms (“sprain” vs “strain”)
- Acronyms
- Copying and pasting code description in the free text field

Analysis Issues

- Sparse data
- Many concepts – how to select the predictive concepts?

In the future of text mining

Analysts will have to decide how to best:

- prepare textual data
- set text mining parameters - synonym dictionaries, word stemming and word combinations.
- optimise the process of utilising textual information
- implement the result of text-mining within their organisations' operational analytical framework.

Thank you for listening!

Questions?

For an electronic copy of the presentation please email
inna.kolyshkina@au.pwc.com