Text Mining for Insurance Claim Cost Prediction

Inna Kolyshkina and Marcel van Rooyen

This project was done in PricewaterhouseCoopers, Sydney, Australia
Phone: +61-2-8266-1429, Fax: +61-2-8286-1429
inna.kolyshkina@au.pwc.com, marcel.van.rooyen@au.pwc.com

Abstract. The paper presents the findings of an industry-based study in the utility of text mining. The purpose of the study was to evaluate the impact of textual information in claims cost prediction. The industrial research setting was a large Australian insurance company. The data mining methodologies used in this research included text mining, and the application of the results from the text mining in subsequent predictive data mining models. The researchers used software of the leading commercial vendors. The research found commercially interesting utility in textual information for claim cost prediction, and also identified new risk management factors.

Keywords: text mining, predictive model, insurance claim prediction, risk management.

1 Introduction

Claims cost prediction is an important focus area for insurance companies. The reason is that proactive case management can significantly reduce the final claims pay-out value. The issue is particularly relevant, when considered that a small number of cases amount to a disproportionately big portion of total claims pay-out value. It follows that small improvements in claims pay-out value prediction, may bring significant financial benefits to insurance companies.

There has been recognition for some time now that data about incidents contain information which allows for a proactive risk management approach (Feyer and Williamson 1998, p.1). The large size of insurance databases is making data mining an increasingly attractive tool for analysis compared to traditional analytical methods (Kolyshkina, Steinberg et al. 2003, p.493). Up to 80% of this data is in unstructured textual format (Feldman 2003, p.481). Realising the potential value of information resident in this textual data, there is growing interest by insurers in the application of new text mining techniques (Feyer, Stout et al. 2001). We refer to an example where text mining analysis of narrative fields about claims, resulted in beneficial claims management and fraud detection in the occupational injury insurance domain (Stout 1998). In the example, the benefits stemmed from information in the textual narrative data, which was not present in the existing coding system.

This paper shows how textual data can be directly included in the claim analysis and used to improve prediction of pay-out value of insurance claims.

G.J. Williams and S.J. Simoff (Eds.): Data Mining, LNAI 3755, pp. 192 – 202, 2006.

The data for the project was provided by a large Australian insurance company. That insurer first wanted to assess the potential value that using text mining facilities could add to the organisation in increasing the precision of claim cost prediction; second to explore the possibilities and benefits of augmenting their existing incident coding system using free text; and thirdly to suggest how text mining could be used for improvement in other areas of the business.

Our approach was to create a model identifying at the time of the incident report, whether the incident would result in a claim pay-out value within the top 10 percent by value, by the end of the next quarter. We assessed the model in terms of the predictive power of textual information on its own, and in terms of textual information adding predictive power to other, non-textual predictors. Predictive power of a model was measured by the cumulative lift the model achieved.

2 Description of Algorithms Used

The researchers used both SAS® Enterprise Miner and SPSS® Clementine text mining software for textual data preparation and text mining. The discovered concepts were similar irrespective of the software package used.

Predictive feature selection and predictive modelling was done using both CART® and TreeNet® (Hastie, Tibshirani et al. 2001). The choice of TreeNet® was because of its high precision in predictive modelling, its effectiveness in selecting predictors, and its resistance to overtraining. CART® models are more easily interpretable than TreeNet® models, therefore the researchers used CART® in conjunction with TreeNet® to assure understandability about the predictive models.

3 Data Description

The first group of data comprised features about claimant demographics, claims pay-out value information, and codings about various aspects of the incident (e.g. about the body part injured). To facilitate the discussion, we name this first group of data TransData. The second group of data which we will name TextData, contained unstructured free-type text fields of about 200 characters each. These fields described the incident and the resulting injury.

Both data groups were identified by claim numbers. The data sets represented all claims reported between 30 September 2002 and 31 March 2004, which were still open at the time of the research. This was an 18 month data history, which provided approximately 56,000 records. The target variable for prediction was a binary indicator (yes/no) of whether or not that injury report had resulted in a claim pay-out value within the top 10 percent by the end of the quarter of the report. The quarter represents a three-month time window of investigation.

4 Description of Analytical Techniques

In Figure 1 we present the research process flow:

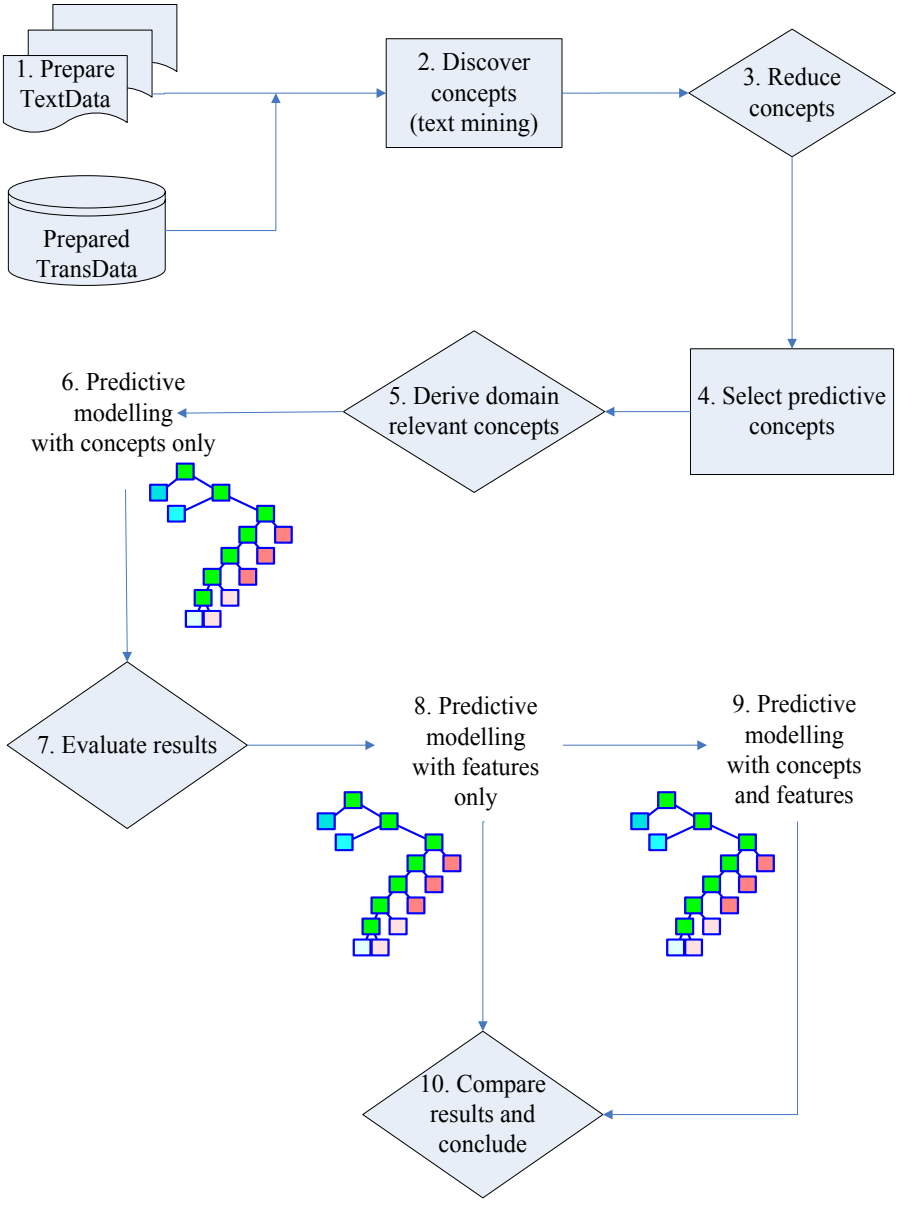

Fig. 1. Research process flow

We will discuss our research process under the following headings. These headings are the same as the labels of the elements of Figure 1.

3.1 Prepare TextData

The process started at point one with the preparation of TextData for text mining. The preparation included routine activities of extracting data, and merging TextData and TransData by the unique claim identity.

3.2 Discover Concepts

We will refer to meaningful words or word combinations resident in the text as "concepts". Text mining is the process of derivation of concepts from unstructured text. This is achieved by applying text mining software to the fields containing free text, where text field is the input and concepts are the outputs. For example, the incident description "the worker fell off horse and broke his leg" may contain such concepts as "worker", "fell off horse", "broken leg".

A concept then leads to creation of the corresponding concept counting variable that assumes the value of n where n is the number of times that the concept is present in the text field. Within this context, text mining can be considered a way of pre-processing unstructured data for use in subsequent modelling. We refer to these concept counting variables from here on as "concept variables".

TextData was mined at point two to derive concept variables. We achieved this by applying SAS ® and SPSS ® text mining software to the data. The mining process was characterised by iterative experimentation to find optimal algorithm settings. These settings included both language and mathematical weightings.

The mining of TextData required not only expertise in the software packages used but also incorporating the subject matter knowledge of the insurance domain.

3.3 Reduce Concepts

About 8000 concepts were discovered in the preceding activity, resulting in a similar number of concept variables. Not only did it prove difficult to make sense of so many concepts, but it would also make subsequent analysis intractable. Further, concepts with a low frequency would not be relevant within our context. Therefore at point three the researchers filtered out those concepts which had a frequency of less than 50 in TextData. After filtering 860 concept variables remained. The issue of the predictive value of concepts now needed resolution. We had prior knowledge about the predictive value of the data features in TransData from existing modelling by the insurer.

3.4 Select Predictive Concepts

The researchers resolved the issue of concept predictability by using TreeNet® at point four to identify the most predictive of the 860 concept variables. We present the nine most predictive concept variables from this step in Table 1. The first column of Table 1 lists the concept variables which were selected by TreeNet®, and the second column states each concept variable's relative predictive importance.

Table 1. Concept importance using TreeNet®

Concept name	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
...	...

3.5 Derive Domain-Relevant Concepts

The researchers depended on insurance domain expertise for deriving additional features at point five. This encompassed the grouping and combining of concepts e.g. 'to injure' and 'to hurt' were set as equivalent terms. Concept derivation was assisted by referring to other targets in addition to the predicted target.

3.6 Predictive Modelling with Concepts Only

At point six the researchers built one TreeNet® and one CART® predictive model for claims cost, using only the concept variables as predictors. The purpose of building the models was to discover the predictive potential of the text-derived concepts. TreeNet® models are known to be highly predictive, but difficult to interpret. Decision trees offer more interpretable results in the form of easily understandable split rules. We therefore also built a CART® decision tree model for improving the interpretability of results for the business.

3.7 Evaluate Results

At point seven the researchers evaluated these two models based on the concepts alone by referring to model topology, gains charts, and model accuracy. The TreeNet® model was 75.7% precise on test data. We present the gains charts for the two models in Figure 2:

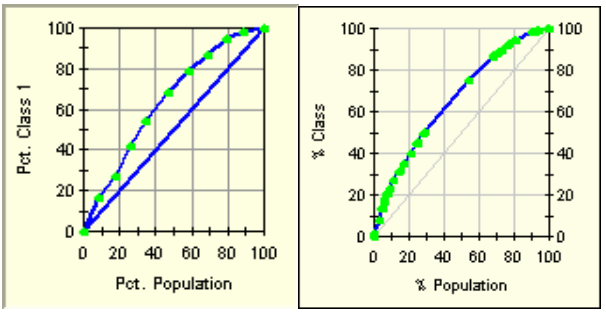

Fig. 2. Gains chart TreeNet® (left)and CART® (right)

From these CART® and TreeNet® gains charts we evaluated that the concepts had predictive potential. We show the reduced model topology from CART® in Figure 3:

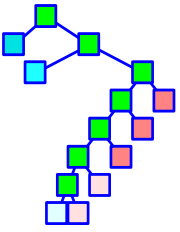

Fig. 3. Topology CART® model using concepts only

CART® tree diagram above is easy to interpret. The first split is on the concept 'eye', followed by 'lacerated', 'stress', 'fracture', 'hernia', 'truck', and 'ladder'. Blue colour indicates lower claim cost while red colour indicates higher claim cost. This topology made domain sense, since eye-related claims have a relatively low cost, as do laceration-related claims.

Table 2. Concept importance using CART®

Concept name	Concept importance
GROUP_EYE	100
GROUP_LACERATED	72.61
FOREIGN BODY	60.85
GROUP_STRESS	60.25
GROUP_FRACTURE	50.45
HERNIA	20.91
KNIFE	14.43
GROUP_LADDER	12.93

In Table 2 we present the most important concepts as identified by this CART® model. The concept names which start with GROUP_, are derived concepts from step five.

3.8 Predictive Modelling with Features Only – Creating a Reference Base

The question of interest for the client was whether adding textual data to the model can improve prediction of claim cost based on existing incident codes. Further, domain experts suggested that textual information may be particularly useful for improving prediction precision where there is high variance in claim pay-out value. This last class of cases was presented in our data by certain injury codes.

To prove that text adds value to the codes, we built two models using only TransData features i.e. there were no concepts in these models. We present the results from these models in section 3.10, where we also compare them to the models which we build at point 9.

The first model we build at point 8 was a CART® model, and we called that model C1. C1 was build using data from all injury codes, in order to maintain understandability for the client across all codes. We built the second model using TreeNet®, and called it called T1. T1 was built using only observations which represented high variance claims pay-out injury codes. This was do demonstrate the potency of concepts in improving predictability in on a high-variance problem.

3.9 Compare Results and Conclude that Concepts Improve Predictive Accuracy

At point nine we build one more CART® model and one more TreeNet® model, also predicting claims pay-out. In these two models we combined TransData features with predictive concept variables from TextData. The combined CART® model (called C2) we built on the same data sets as C1 above. The combined TreeNet® model (called T2) we build on the same data sets as T1 above.

In the next section we will compare the results from these four models, and make inference about the contribution of the concepts to predictive accuracy about claims cost.

3.10 Compare Results and Conclude

We present the gains charts of the C1 and C2 models in Figure 4:

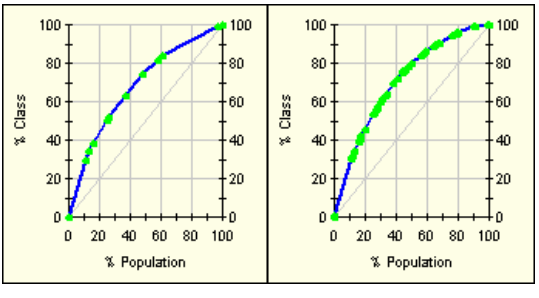

Fig. 4. Gains chart CART® models – C1 left and C2 right

Comparing the two gains charts shows about a 5% increase in predictive performance of C2 over C1. In Figure 5 we present the gains charts of TreeNet® models T1 and T2:

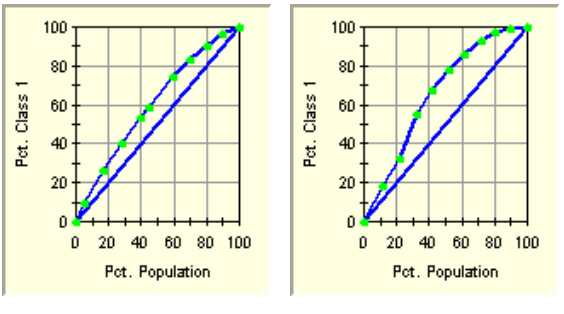

Fig. 5. Gains charts TreeNet® – T1 left and T2 right

The gains chart of T2 shows about a 10% improvement over the gains chart of T1. We note that this improvement is twice as much as the improvement over the full data set (C2 over C1). This is consistent with the idea that textual features add value to claims pay-out prediction in high-variance injury code categories.

We attribute the increase in predictive value of the combined models, to resolution which is added to the models by the addition of the concepts. In Figure 6 we display the additional resolution in the C2 CART® model.

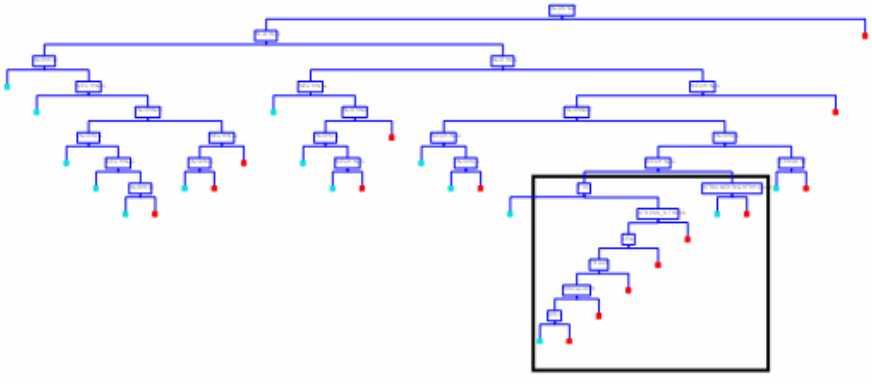

Fig. 6. Additional CART® model resolution from concepts

The binary splits outside the boxed area in Figure 6, are on features from TransData e.g. injury location (part of body), nature of the injury, and mechanism of the injury. The splits inside the square were on concepts from TextData. The concepts found in the square are not covered by the insurer's existing coding system. Examples of such concepts were 'truck', 'roof', and 'ladder'. Such concepts can also be useful in identifying particular industry OH&S issues, which require proactive risk management strategies.

From the above results, the researchers conclude that the combined models had increased predictive precision in claim cost prediction, compared to the models with TransData features only. The conclusion is valid for the full data set, as well as for the high variance data subset.

Mining unstructured textual data therefore did improve the prediction of insurance claims cost. We also discovered new concepts which complement the insurer's existing coding system. The next section contains ideas about how text mining could be used for improving other areas of the insurer's business.

4 Potential Applications of Text Mining by the Insurance Industry

The researchers offer their impressions of some possible uses of text mining in the insurance industry;

- first, we perceive application in an insurer's capital management. Improved prediction about future insurance claims cost, should bring about improvements in the quantification of re-insurance needs. Such quantification about needs will be invaluable in negotiating favourable re-insurance premiums, and terms and conditions, with re-insurers. Further, such quantification should assist with better planning about working capital requirements for that portion of their risk which is not re-insurable;
- a second potential application is where textual data from the incident investigation is available for analysis. Insurers could identify new risk factors, and previously unknown interactions between known risk factors. Such knowledge would enable insurers to develop industry-specific proactive risk management practices with their insured clients, with consequential financial benefits to both parties;
- a third potential application is where textual data is available from after-the-event therapy sessions with victims. Here, text mining could be used to discover concepts about victims' attitude about and perception of risk, leading up to the risk event (Dedobbeleer and Béland 1998). Knowing such psychographic factors will enable insurers to develop industry-specific, proactive behaviour management programs with their clients. Such programs could bring about paradigm shift in the approach to risk management;
- a further potential application is quality control of the application of incident codlings in reporting systems, and for development of new incident reporting and investigation business rules;
- further, in those cases where textual data is available from market research interviews with potential retail insurance customers, text mining can be used to discover consumers' attitudes about, and needs for insurance. Such concepts can then be used for segmenting the retail market simultaneously for attitude and need, using conceptual clustering techniques. Such multi-dimensional market segmentation would enable insurance retailers to develop campaign offers which are much better matched to consumer needs and attitudes, than what is possible, for instance, with demographic segmentation approaches. Such campaign offers will result in better campaign response rates, and improved competitive advantage for insurers (Berry and Linoff 2004) Chapter 4;
- the use of the data mining techniques has been shown to be effective in fraud detection and prevention (Phua, Lee et al. Submitted). Text mining is already being used in the insurance claims fraud discovery and prevention arena (Mailvaganam Accessed February 2005) (Ellingworth and Sullivan Accessed February 2005), and the researchers perceive an underutilised opportunity by the insurance industry. This entails mining unstructured textual data from claimants' contact with the insurer. Such mining would be focused at discovering both known and new concepts from that data, upon which to base both proactive and remedial fraud management.

5 Future Work

The researchers identify a number of issues which require further research. The first is investigating the value of text mining in predicting other targets of interest to the

insurance industry. Such other targets could be 'number of days off work', 'cost of medical treatment', or 'exact claim pay-out amount'.

Further, research could extend the modelling time horizon about the target past the current three-month period, to for instance six months or even 12 months. This will enable insurers to accordingly extend their planning horizon about claims cost. Research about the utility of text mining for business rule development and fraud management could be valuable.

Research is also required into some of the opportunities identified in the previous section to:

- quantify the dollar benefits from improved capital management which is realisable from text mining;
- discover new risk factors or interactions between risk factors;
- use text mining as a psychographic profiling and segmenting tool; and
- quantify the monetary benefits from marketing campaigns based upon this psychographic approach compared to traditional marketing approaches.

Acknowledgements

The researchers acknowledge the contributions to this research of Michael Playford and Anne-Marie Feyer, both PricewaterhouseCoopers partners, for offering help and valuable insights. Dominic Roe, consultant PricewaterhouseCoopers, and Bianca Zubac, student UNSW for their assistance in project facilitation and data analysis. SAS Institute and SPSS, both for providing text mining software for the project as well as for assistance and guidance in using text mining software.

References

Berry, M. J. A. and G. S. Linoff (2004). Data Mining Techniques for Marketing, Sales, and Customer Relationship Management. Indianapolis, Wiley.

Dedobbeleer, N. and F. Béland (1998). Is risk perception one of the dimensions of safety climate? Occupational Injury: Risk, Prevention and Intervention. A.-M. Feyer and A. Williamson. London, Taylor & Francis**:** 73-81.

Ellingworth, M. and D. Sullivan (Accessed February 2005). Text Mining Improves Business Intelligence and Predictive Modelling in Insurance, http://www.dwreview.com/article_sub.cfm?articleId=6995.html. **2005**.

Feldman, R. (2003). Mining Text Data. The Handbook of Data Mining. N. Ye. London, Lawrence Erlbaum Associates**:** 481-517.

Feyer, A.-M., N. Stout, et al. (2001). "Use of narrative analysis for comparisons of the causes of fatal accidents in three countries: New Zealand, Australia, and the United States." Injury Prevention **7**: i15-i20.

Feyer, A.-M. and A. Williamson (1998). Introduction. Occupational Injury: Risk, Prevention and Intervention. A.-M. Feyer and A. Williamson. London, Taylor & Francis**:** 1-3.

Hastie, T., R. Tibshirani, et al. (2001). The elements of statistical learning: Data mining, inference and prediction. New York, Springer-Verlag.

Kolyshkina, I., D. Steinberg, et al. (2003). Using Data Mining for Modeling Insurance Risk and Comparison of Data Mining and Linear Modeling Approaches. Intelligent and Other Computational Techniques in Insurance: Theory and Applications. A. F. Shapiro and L. C. Jain. London, World Scientific. **Volume 6:** 493-421.

Mailvaganam, H. (Accessed February 2005). Text Mining for Fraud Detection: Creating cost effective data mining solutions for fraud analysis, http://www.dwreview.com/Data_mining/Effective_Text_Mining.html . **2005**.

Phua, C., V. Lee, et al. (Submitted). "A Comprehensive Survey of Data Mining-based Fraud Detection Research." Submitted.

Stout, N. (1998). Analysis of narrative text fields in occupational injury data. Occupational Injury: Risk, Prevention and Intervention. A.-M. Feyer and A. Williamson. London, Taylor & Francis**:** 15-20.