Paper 071-30

Predicting Workers' Compensation Insurance Fraud Using SAS® Enterprise Miner™ 5.1 and SAS® Text Miner

Terry J. Woodfield, SAS Institute Inc., Irvine, CA

ABSTRACT

Insurance fraud costs the property and casualty insurance industry over 25 billion dollars (USD) annually. This paper addresses workers' compensation claim fraud. A data mining approach is adopted, and issues of data preparation are discussed. The focus is on building predictive models to score an open claim for a propensity to be fraudulent. A key component to modeling is the use of textual data to enhance predictive accuracy. Binary response modeling is emphasized, but a two-stage modeling approach is briefly addressed. SAS Enterprise Miner 5.1 with SAS Text Miner provides the modeling environment for the analysis.

INTRODUCTION

Insurance fraud costs the property and casualty insurance industry over 25 billion dollars (USD) annually. Workers' compensation insurance alone accounts for a sizable portion of this total cost. Workers' compensation fraud can be divided into three types:

- **Claimant fraud** occurs when an individual falsely claims to have experienced a work-related injury or exaggerates a legitimate injury and files for insurance benefits.
- **Provider fraud** occurs when a provider bills an insurance company for services that were not provided.
- **Internal fraud** occurs when someone inside the insurance organization creates a fictitious claim to route benefits to an accomplice.

Sometimes two or more individuals conspire to commit fraud, which leads to fraud cases that have elements of two or three of the possible types of fraud. For example, a claimant and a provider might conspire to commit fraud.

A standard operational practice is to have claims adjustors or claims supervisors identify potential fraud cases and route them to a Special Investigative Unit (SIU). The "traditional" SIU employs professionals who have skills in fraud investigation such as surveillance, evidence collection, and evidence analysis. These SIU investigators often have a law enforcement background. One shortcoming of the traditional SIU organization is the absence of computer data processing and predictive modeling professionals. Consequently, details of fraud investigations are rarely stored in electronic form in a modern database.

This paper outlines a strategy for detecting claimant fraud in workers' compensation insurance. It is assumed that a traditional referral process and SIU organization exist. Suggestions are made to help an organization transition to a modern data mining approach to fraud detection. While the primary impact will be on the referral process, there are implications for improved data processing to facilitate future predictive modeling.

INSURANCE FRAUD CONTRASTED WITH OTHER FORMS OF FRAUD

Modern success stories for data-mining fraud detection include applications in credit-card fraud and telecommunications fraud, which are much easier to model than insurance fraud. Credit-card and telecommunications fraud is easily identified within a relatively short period of time because a customer contests a bill or reports a stolen credit card. Essentially, the absence of a payment is a key component of fraud in credit and telecommunications applications.

With insurance fraud, if the fraudulent behavior is not discovered, the insurer never knows that the fraud has occurred. Consequently, an uninvestigated claim cannot be labeled with respect to fraud. Only claims that are investigated by the SIU can be labeled FRAUD=YES or FRAUD=NO. This implies that, out of hundreds of thousands of claims that have detailed claim information, only about 1% can be used to train a predictive model. Contrast this to credit-card fraud where all cases can be included in modeling. Part of a proposed solution to the insurance fraud problem is to process the uninvestigated insurance claims so that some of these cases can also be used in modeling.

THE DATA MINING PROBLEM

The insurance fraud problem translates to two data mining problems.

1. Given a set of records (claims) that do not have known target values, use unsupervised learning techniques to divide the data into two or more clusters, and employ domain expertise to evaluate each cluster as likely to be either FRAUD=YES or FRAUD=NO. This can be an iterative process that involves modifying unsupervised learning options and criteria until domain experts are satisfied with the clusters that are produced.

2. Given a set of records (claims) that have known or estimated target values, construct a predictive model to score new cases with respect to a propensity for being fraudulent.

DATA PREPARATION

Table 1 lists the fields (variables) that are common in a workers' compensation database.

Field	Description
Claim ID	Primary index for extracting claimant, policy, and transactional data from the database
Birth Date	Claimant birth date
Injury Date	Date of injury
Employment Location	Location of employment (ZIP code)
Accident Location	Location of accident (ZIP code)
Gender	F/M
Body-Part Code	Part of body injured
Accident Code	Accident code (slip, fall, etc.)
Nature Code	Nature of injury (laceration, contusion, etc.)
Industry Code	Industry listed on insured company's policy
Occupation Code	Occupation
Prior Injury	Y/N for related prior injury
Risk Factors	Multiple Y/N fields for risk factors such as anemia, diabetes, etc.
ICD9 or CPT Codes	Standard medical procedure, disease, and condition codes, usually stored in multiple fields for primary, secondary, etc., codes
Medical Codes	Multiple Y/N fields for medical fields such as ER, inpatient hospital, outpatient hospital, radiology, physical therapy, etc.
Rehab Codes	Multiple rehabilitation fields such as vocational rehabilitation, education, and re-training
Adjustor Notes	Free-format text field with adjustor notes about the case

Table 1. Common Variables in a Workers' Compensation Database

Additional fields are created from claim transactions. For example, some claim systems have a field in a claims master table for emergency room (ER) visits, while other systems set the ER flag dynamically using hospital payment transactions. The existence of a positive payment with an ER payment code implies ER=YES.

A sufficiently detailed insurance database allows a predictive modeler to derive model inputs that have a conceptual basis for indicating fraud. Here are some examples:

- The distance between claimant's home address and a medical provider's office
- Statistics for transactions by provider
- Changing providers for the same treatment correlated with other claim activity.

The first example might raise questions about the distance a claimant must travel to a provider when similar providers are closer. The second example is related to a common fraud practice, that is, a lawyer recruits a claimant and sends the

claimant to a chiropractor for 10 closely spaced visits. The lawyer then contacts the insurance company and seeks a settlement based on the existing costs and the assertion that many more visits to the chiropractor will be necessary because of pain from the injury. High frequency and low variability can indicate this type of fraud. The third example might provide more evidence of the fraud type described in the second example, if a change in provider (for example, from a physical therapist to a chiropractor) occurs at about the same time that the insurance company is notified of lawyer representation, then it is possible that the claimant is conspiring with the lawyer and the chiropractor to exaggerate injuries.

Measuring distances based on addresses is a common data preparation activity. The SAS Course Notes for Data Preparation (SAS Institute Inc., 2004a) provide macros for finding the distance between two ZIP-code regions. The primary calculation involves using the Haversine function, which gives the distance between two locations that are specified as latitude and longitude coordinates.

The unsupervised learning techniques, which are used later in this paper, require that all inputs be measured on an interval scale. Consequently, categorical inputs cannot be used without some form of transformation. The usual approach of dummy coding does not work well for unsupervised learning. The Data Preparation course notes describe two techniques for converting categorical data to numeric data. These are the weight-of-evidence approach and the smoothed-weight–of-evidence approach. Georges (2004) provides empirical evidence to support the smoothed-weight-of-evidence approach for predictive modeling. For a complete description of the technique, see Georges (2004) or SAS Institute Inc. (2004a) Smoothed-weight-of-evidence coding requires knowledge of a target variable. Therefore, this approach cannot be used in the preliminary unsupervised learning step. You should avoid using categorical variables in unsupervised learning.

UNSUPERVISED LEARNING TECHNIQUES FOR FRAUD DETECTION

If SIU data on fraud cases is unavailable, you can still derive a fraud model. Using numeric inputs, group the data into clusters by using unsupervised learning techniques. Because no target is available, you can either substitute domain expert weights for categories, or you can omit categorical variables. The safest approach seems to be to omit categorical variables from unsupervised learning.

SAS Enterprise Miner 5 supports k-means clustering, self-organizing maps (SOM), and Kohonen vector quantization for clustering data. By default, a maximum of 40 clusters are allowed, and the cubic clustering criterion is used to deduce a “best” number of clusters. Using this default setting, you can derive clusters for the data. Domain experts then review the clusters and determine if any clusters seem indicative of fraud. The domain expert must have access to the variables for each case in a cluster. The smallest clusters should be examined first.

Ideally, your organization will have domain experts who have some experience in fraud cases. Because the task of examining each case in each cluster can be overwhelming, you should have the domain experts describe cases that imply fraud. For example, flag cases where the primary medical provider is located more than 50 miles from the claimant’s home, then calculate the proportion of cases that have this flag set to YES for each cluster. Obviously, you should flag cases that have legal representation. If you can describe five or more rules that might indicate fraud, you should be able to automate screening without having to examine every case.

Text data is a special kind of categorical data in which the cardinality is essentially infinite. However, text mining techniques like those supported by SAS Text Miner convert text data into a set of interval scaled numeric values. For a complete description of text mining using SAS Text Miner software, see SAS Institute Inc. (2004b). The Adjustor Notes are treated as documents in a collection by SAS Text Miner. A variety of options are available for determining the importance of terms in a document and assigning weights to these terms. One document will be transformed into one or more interval scaled variables, and these variables may be used in unsupervised learning.

The goal of unsupervised learning is to derive clusters in which one or more clusters have a high proportion of possible fraud cases as diagnosed by domain experts, and the remaining clusters have a low proportion of possible fraud cases. The clustering score code is used to score all cases as FRAUD=YES or FRAUD=NO based on the cluster in which an observation falls. You will be fortunate if you obtain this goal. You can modify options or try different methods to obtain adequate clusters. There are many choices.

- Try alternative algorithms. For example, if k-means fails, try Kohonen vector quantization.
- SOM methods allow specification of a grid that can be varied to try to get better results.
- Different distance measures can be used in cluster data..
- If you have text variables, try different options in SAS Text Miner. Consider using both sets of numeric variables that are produced by SAS Text Miner. These are singular value decomposition variables and roll-up term variables.

If you fail to obtain adequate clusters, then you must code each case, manually. While manually coded targets can be superior to targets that are defined by clusters, manual screening can be impractical. Furthermore, fatigue and human bias can have a negative impact on manual target coding.

SUPERVISED LEARNING TECHNIQUES FOR FRAUD DETECTION

Supervised learning techniques can be performed by using manually coded or cluster coded (pseudo) targets, or by using actual targets from the SIU. Realistically, any case that is not investigated by the SIU is an "I don't know" case, so manual coding or cluster coding will be required to take advantage of all the data. You should not treat cases that were not investigated as FRAUD=NO cases.

Manual coding or cluster coding provides a rough approximation that often leads to successful models. However, when models are used to screen cases for referral to the SIU, then low-scoring cases might be false negatives. Ideally, you can randomly sample low-scoring cases and have these investigated by the SIU along with all the high-scoring cases. Then you can use these results to estimate the proportion of false negatives in the low-scoring cases. This information can be used to improve future models.

SCORING NEW CASES

SAS Enterprise Miner can produce score code in the SAS language, in ANSI standard C, and in Java. If the SAS language is used, new cases can be scored by using Base SAS. However, if text mining is used to quantify textual data, then only sites that license SAS Text Miner can score new cases, because SAS Text Miner score code calls external executables that exist only on systems that have SAS Text Miner licensed.

EXAMPLE

Data is available for workers' compensation cases that cover a five-year period. Included in the data is the text field Adjustor Notes, so SAS Text Miner will be used in the analysis. The target values are mostly manually coded because the data was obtained for a pilot study. Only a few cases were available from the SIU.

Figure 1 shows a typical flow diagram for an insurance fraud project.

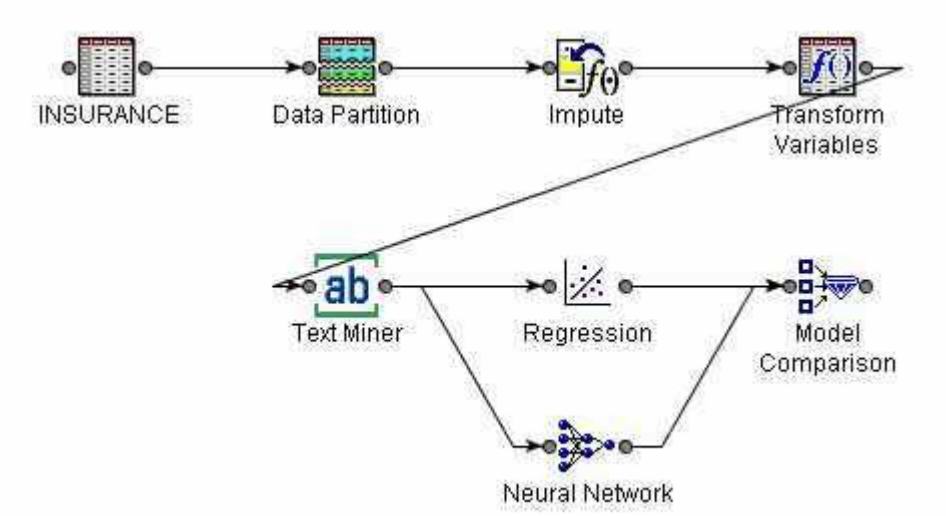

Figure 1. Insurance Fraud Project Flow Diagram

Figure 2 shows a lift chart that compares the logistic regression model to the neural network model derived as shown in Figure 1. You can see that both models are adequate for prediction. Comparison of models with and without text mining inputs reveals that text mining substantially improves the accuracy of the model.

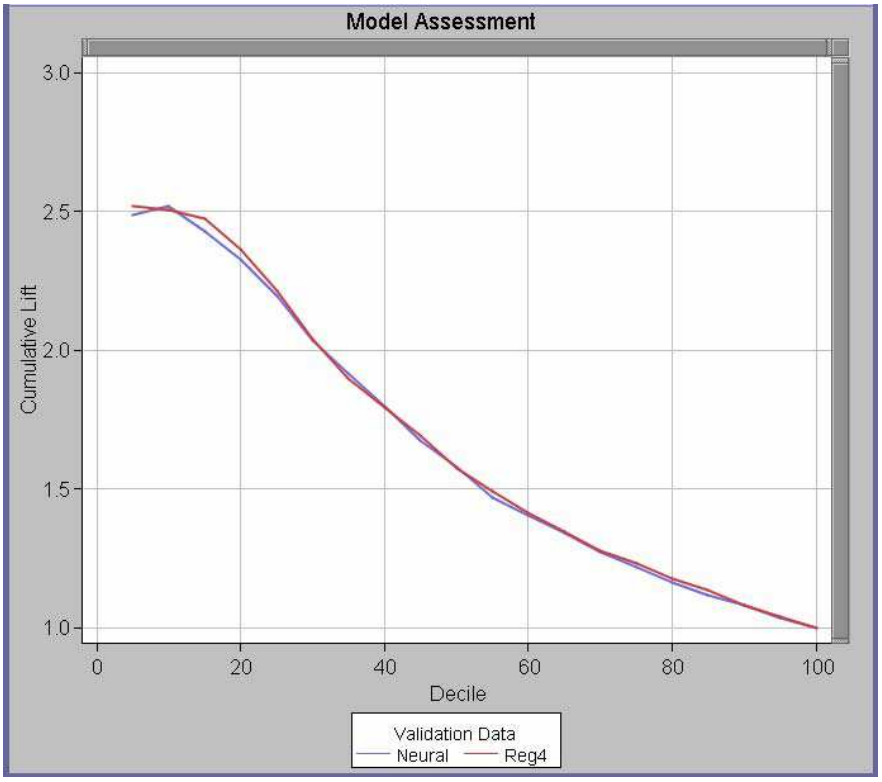

Figure 2. Lift Chart Comparing Two Models

A TWO-STAGE FRAUD MODEL

Based on several criteria, the SIU decides to investigate a referred claim. One criterion is the expected loss due to fraud. Because a typical investigation costs more than 1,000 dollars (USD), if fraud prevention or recovery is estimated to be less than 1,000 dollars, then it is not worth the effort to pursue the case. (If fraudsters realize that they can get 1,000 dollars with no risk and low-cost claims become excessive, then this policy will change.) When a new case is referred to the SIU, the loss estimate is entered in the claim system so that the SIU can look at this amount and decide whether to pursue the case. If, instead of adjustor loss estimates, a predictive model is used to estimate loss, then a two-stage fraud model can be developed. Cases that have low loss amounts are unlikely to get high scores for referral.

Forecasting loss in insurance requires the use of censored prediction models, also called survival models. Speights, et al (1999), describe a neural network survival model for predicting claim duration. Duration times compensation rate produces a prediction for wage losses. Predicting medical costs and costs for vocational rehabilitation is more difficult.

CONCLUSION

Fraud modeling relies heavily on good data and domain expertise. Usually, the addition of text mining tools results in increased accuracy. Insurance fraud modeling is very different from many fraud modeling situations because cases that have not been flagged FRAUD=YES as a result of an investigation are really FRAUD=MAYBE rather than FRAUD=NO.

REFERENCES

Georges, Jim (2004). “Using non-numeric data in parametric prediction,” M2004: Seventh Annual Data Mining Conference, Las Vegas, Nevada.

SAS Institute Inc. (2004a), Data Preparation for Data Mining Using SAS Software Course Notes.

SAS Institute Inc. (2004b), Mining Textual Data Using SAS Text Miner for SAS9 Course Notes.

Speights, David, Brodsky, Joel, and Chudova, Darya (1999). “Using Neural Networks to Predict Claim Duration in the Presence of Right Censoring and Covariates,” Casualty Actuarial Society Forum, 255-278.

ACKNOWLEDGMENTS

Jim Georges created the data preparation course notes (SAS, 2004) and wrote the programs for distance measures, transaction processing, and categorical data transformations.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author:

Terry Woodfield
SAS Institute Inc.
5 Park Plaza, Suite 900
Irvine, CA 92614
Work Phone: 949.852.8550x321
Fax: 949.852.5277

Email: terry.woodfield@sas.com
Web: www.sas.com