Paper 171-2010

Improving the Defense Lines: The Future of Fraud Detection in the Insurance Industry (with Fraud Risk Models, Text Mining, and Social Networks)

Terisa Roberts, SAS Institute Inc., Marlow, United Kingdom

ABSTRACT

Given the current global economic turmoil and contracting economies, financial crime is on the rise. The use of analytical techniques to protect financial institutions against fraudulent activity has seen varying degrees of success in the past. Recent advances include the use of rule-based fraud detection flags, exception reporting, third-party data searching, profiling, and fraud scorecards based on quantitative data. More recently, advanced analytical techniques such as text mining and social networks have also been used to effectively support the fraud investigation process. Artificial intelligence algorithms can be used to detect human involvement where it is not expected, even where suspicious activity has not yet been detected. This paper will look at a comprehensive framework to combine the results from text data analysis, social networks, and artificial intelligence in order to improve the accuracy of fraud risk models, while also maintaining an easy-to-implement and easy-to-interpret design.

The paper will include the results of its application using data of a major insurance company in South Africa.

INTRODUCTION

Despite recent positive developments, the cost of fraud to the insurance industry continues to rise. Fraud is big business today. According to the Insurance Fraud Bureau, fraud adds 5% to the average insurance premium in the UK. In other countries, such as South Africa, it is estimated to add as much as 15% to the average premium. However, these are estimates because it is virtually impossible to determine an exact value for the amount of money that is stolen through insurance fraud. More alarmingly, in the United States, new data from the National Insurance Crime Bureau uncovered an increase in the number of questionable claims that are related to cases of insurance fraud in recent times as the economy continued its downward spiral (Risk & Insurance, 2009). The increase in fraudulent activity is echoed in the press in the United Kingdom, where Allianz Insurance reported that fraudulent claims have doubled in the first three months of 2009 as firms struggle to keep their heads above water in the current recession. By their natures, insurance fraud crimes are designed to be undetectable, which means that a significant amount of fraudulent activity still goes unnoticed. And, as has become all too evident in recent years, a single fraud can wipe out years of profit, drive away investors, ruin a brand, or bankrupt even the largest organization. The 2007 collapse of Independent Insurance in the UK is an example of what can happen.

The advances in technology and the growing pressure on insurance companies to serve consumers through their traditional direct channels, as well as newer distribution channels like the Web, are opening up new opportunities for fraud.

Insurance companies realize the importance of combating fraud; however, the following major challenges still exist:

- Companies investigate suspicious activity after-the-fact, rather than pro actively. According to several industry sources and players this needs to change.
- A tension exists between the need to maximize profits on the one hand and to invest in anti-fraud measures on the other.
- A lack of resources in specialist fraud investigation teams.
- Once fraud has been identified, very little is being done to use the new-found information to stop the activity from happening again.
- For an analytical data driven solution, known fraud cases are rare and what is currently investigated and reported is known to be only the “tip of the iceberg.”

THE WIDE SPECTRUM OF FRAUD

Fraud might be committed at different stages in the insurance transaction and by different parties: applicants for insurance (new customers), policyholders (existing customers), third-party claimants, and professionals who provide services to claimants. Common frauds include "padding," or inflating actual claims; misrepresenting facts on an insurance application; submitting claims for injuries or damage that never occurred; and "staging" accidents. Those who commit insurance fraud range from organized criminals to professionals and technicians who inflate the cost of services or charge for services that were not rendered, to ordinary people who want to cover their deductible or view

filing a claim as an opportunity to make money.

TRADITIONAL FRAUD DETECTION TECHNIQUES

A summary of anti-fraud measures is outlined below. These measures increase in analytical complexity and data intensity as you move down the list. All can be either supported or implemented using SAS® software. A combination of measures should be implemented to improve the lines of defense against fraud.

- **Whistleblower hotlines**

Hotlines are commonly used by most insurance companies as one of the first lines of defense against fraud.

- **Internal audit procedures**

Internal audit procedures are seen as the second most common method of fraud detection.

- **Watch lists (Internal and 3rd party)**

Matching entities against available internal and external watch lists is effective in identifying organized criminals.

- **Diagnostic fraud indicators**

Based on industry knowledge, diagnostic fraud indicators are used to identify circumstances that suggest greater statistical significance that the case might contain elements of deceit. The claims handler would typically be responsible for completing a survey on the diagnostic fraud indicators before the claim will be processed, so these indicators are usually based on the biased judgment of the claims handler. Unconsolidated data in disparate systems makes it difficult to test these checks and fraudsters are quick to learn the rules.

- **Anomaly detection**

Similar to exception reporting, Anomaly detection refers to detecting patterns in a given data set that do not conform to an established normal behavior. The thresholds can be determined by expert judgment or statistical techniques. The patterns that are detected are called anomalies and are often translated into critical and actionable information. Anomalies are also referred to as outliers, abnormalities, deviations, exceptions, or peculiarities. This is similar to the familiar exception reporting.

- **Profiling**

Profiling is used to construct an outline of an entity's individual characteristics. Profiling can be done by using cluster analysis or segmentation analysis (for example, to compare a new claim with the typical profile of a suspicious claim).

ADVANCED ANALYTICAL FRAUD DETECTION TECHNIQUES

- **Predictive modeling techniques (including decision trees, regression analysis, and neural networks)**

Trained classifications include any predictive modeling technique where a model is fitted on a sample of known fraud cases and known good cases. The system might also output reason codes that indicate relative contributions of various variables to a particular result. Some advanced techniques, such as neural networks, are used, but since fraudsters are quick to alter their behavior, more rules are continuously needed to improve performance of these models.

- **Expert systems**

Expert systems are a type of artificially intelligent system, which stores expertise concerning some subject matter in a knowledge base and attempts to solve problems in a manner that simulates the thought processes of a

human expert. Due to the low frequency of known fraud cases, unsupervised classifications, or a hybrid of both are also used.

- **Text mining**

Text mining is the process that uses a set of algorithms for converting unstructured text into structured data objects. Techniques are available to deal with semantics, syntax, stemming, part of speech tagging, and the identification of entities. Text mining can be seen as an exploratory tool to discover meaningful information that resides in textual data fields like the claim narrative. The quantitative results from the text mining analysis can be incorporated directly into the structured predictive models to improve performance.

- **Social network analysis**

Social network analysis is a study of social relationships in terms of nodes and ties. Nodes are the individual actors/entities within the networks, and ties are the relationships between these actors. These relationships can be strong and obvious (hard links), for example, a married couple sharing the same home address. These relationships can also display soft links, where entities demonstrate similar behavior. With very large networks found in insurance data, it is required to detect communities or sub-networks.

THE CASE FOR A SUSPICIOUS ACTIVITY ASSESSMENT SYSTEM

A comprehensive fraud management strategy should include measures to prevent, deter, recognize, detect, and investigate fraudulent activity. Given the challenges that are outlined above, an effective suspicious activity assessment system should be proactive, accurate, fast, flexible, consistent, and transparent. The system should leverage the wealth and volume of data that are available as well as the expertise and experience of the fraud investigation specialists.

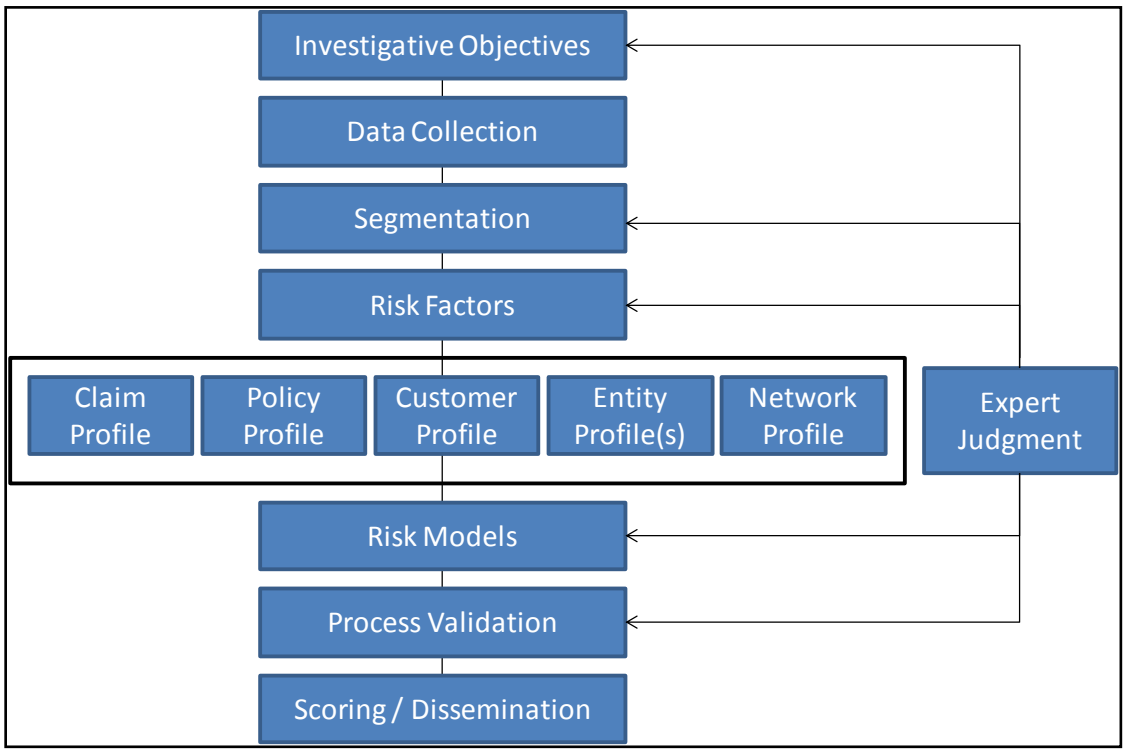

Figure 1. Suspicious Activity Assessment System: Analytical Process

The suspicious activity assessment system is designed to integrate numerous entity-level risk models into a comprehensive architecture, incorporating the expert judgment of fraud investigation specialists in the identification of segmentation rules, risk factors, final model selection, and process validation (as outlined in figure 1). To develop proactive profiles of the interested parties, the system is designed to score entities (customers, brokers, service

providers, and so on) at regular time intervals, including customer acquisition, claim submission, when a customer requests a new policy, and so on.

A hypothetical example of a customer-level scorecard is provided in figure 2. Based on the probability of fraud and estimated potential loss, claims are prioritized for investigation. For example, the probability of fraud is predicted with a logistic regression model and the fraud exposure is predicted by a general linear model. The text mining and social network analysis results are incorporated either directly into the models or weighted based on expert judgment. Business decisions can be made by using the risk matrix, where highly probable and severe cases should receive first priority.

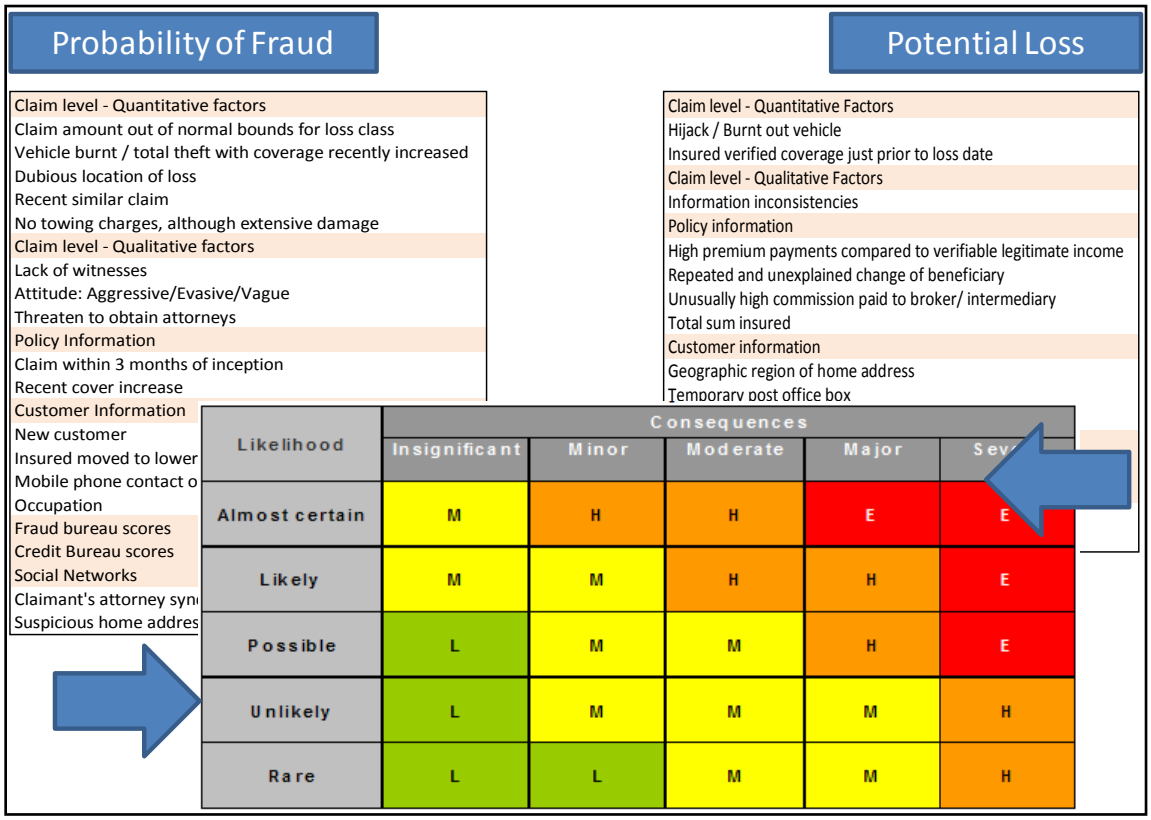

Figure 2. Hypothetical Example of Fraud Risk Scorecards

EXAMPLE USING SAS SOFTWARE

A suspicious activity assessment system can easily be implemented using SAS software including Base SAS® for data processing, SAS® Enterprise Miner™ for advanced analytics (including text mining and fraud scorecards), and SAS® Social Network Analysis software for community detection in large networks and ad hoc queries to support ongoing investigations. In this section, we look at a simplified example of the implementation of a suspicious activity assessment system using the data from a large insurance company.

1. **Investigative objectives**

   During this phase, the most costly and urgent types of crime need to be identified, together with the organizational objectives, business processes, and a better understanding of required preemptive actions.

   Say, the insurance company identified the following organizational objectives:

   - The cost and occurrences of insurance fraud have reached unacceptable levels and need to be lowered.
   - Once suspicious activity has been flagged, investigations take too long due to a lack of a centralized data platform, while a large proportion of investigated cases are unfounded, due to ineffective red flags. The investigation periods need to be shortened.
   - An automated suspicious activity assessment system is required to be implemented due to a lack of resources to check for fraud and verify all insurance claims.

2. **Data collection and pre-processing**

   Data is typically sourced from the data warehouse. Because the data source represents the starting point for higher-level business analytics, data cleansing and data pre-processing efforts should not be underestimated. There are many causes of poor data quality, which need to be addressed.

3. **Segmentation**

   Segmentation rules can be business driven (for example, by market segment or type of loss class) or data driven (using clustering algorithms). Due to the unique properties of fraud cases that are typically flagged for investigation, specific segmentation rules are required to classify the types of fraud cases correctly. Text mining algorithms in SAS Enterprise Miner can be used to analyze unstructured textual data of previously investigated cases to effectively identify accurate segmentation rules and interested and implicated entities.

4. **Identification of risk factors**

   The expert judgment of fraud specialists, accessible fraud detection indicators according to industry knowledge, and the results from an exploratory data analysis can be used to identify a comprehensive set of potential risk factors.

5. **Set up of entity-level fraud scorecards**

   Logistic regression is successfully used for credit scoring because it provides an easy-to-interpret and easy-to-implement model design. Research also suggests that the use of less complex and faster algorithms might produce equal, if not better, results than complex non-linear supervised approaches (Phua, et al, 2002).

   In our example, fraud scorecards (incorporating text mining and network analysis results) are developed on the following entity levels:

   - customer (see example in figure 3)
   - broker
   - service provider

6. **Process validation**

   This phase requires a thorough evaluation of the steps that were executed to construct the models. The entire process is iterative, and the process validation phase should ensure and validate results before final deployment. The models require continuous learning and monitoring to adapt to the ever-changing characteristics of criminal behavior.

7. Dissemination of information

The dynamic dissemination of information is a crucial element in the investigation process. In addition to the flexible business intelligence capabilities of SAS, SAS Social Network Analysis software provides a network visualization interface to present a complete picture of entities, their characteristics, and networks.

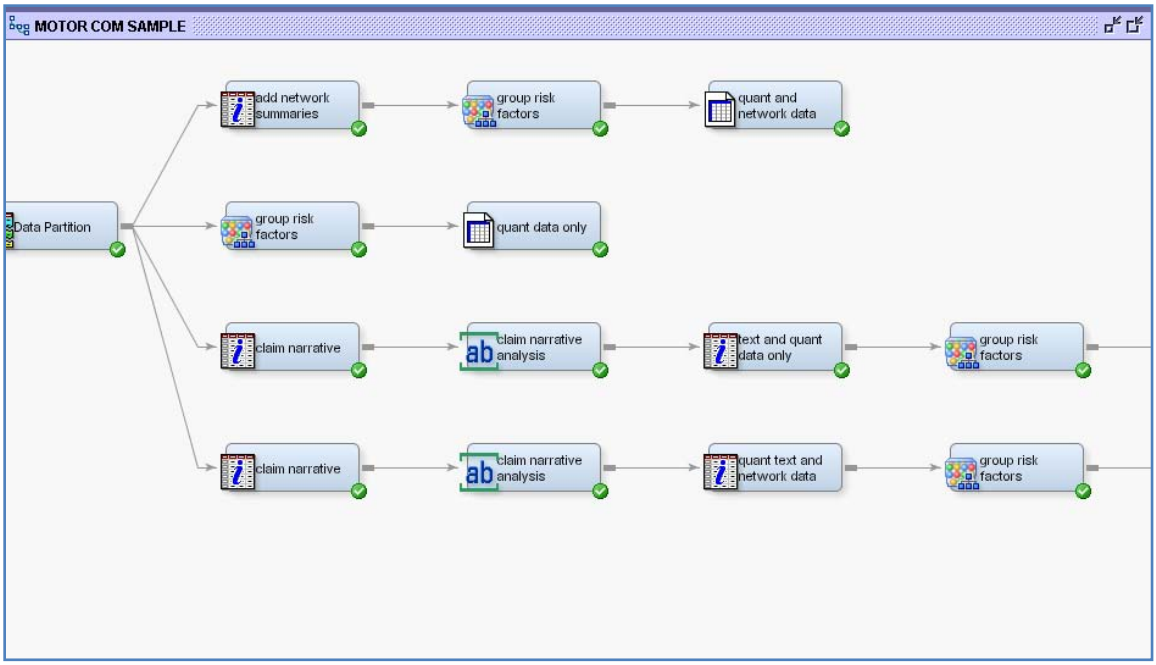

Figure 3. Fraud Risk Scorecards (with Text Data and Network Summaries) in SAS Enterprise Miner

CONCLUSION

Insurance fraud remains a big challenge for the industry, regulatory authorities, and the general public. This paper has demonstrated how the performance of a comprehensive fraud detection system can be greatly enhanced by using a blend of the powerful analytical capabilities of SAS, including text mining algorithms, fraud risk scorecards on multiple entity levels, and network summaries. Such a system enables an insurance company to recognize and detect fraudulent activity more accurately and rapidly, prioritize and improve the quality and quantity of their investigations, reduce their fraud expenditure, and uncover organized crime.

REFERENCES

Abrahams, Clark,and Mingyuan, Zhang. 2009. *Credit Risk Assessment: The New Lending System for Borrowers, Lenders, and Investors*. Hoboken, NJ: John Wiley & Sons, Inc.

Blondel, V. D., et al. 2008. “Fast Unfolding of Communities in Large Networks.” *Journal of Statistical Mechanics: Theory and Experiment*

Bolton, R. J., and D. J. Hand. 2002. “Statistical Fraud Detection: A Review.” *Statistical Science* 17(2): 235-255.

Caudill, S. B., M. Ayuso, and M. Guillén. 2005. “Fraud Detection Using a Multinomial Logit Model with Missing Information.” *Journal of Risk and Insurance* 72(4): 539-550.

Insurance Fraud Bureau. 2010. http://www.insurancefraudbureau.org/.

Lilley, Peter. 2003. *Dirty Dealing: The Untold Truth about Global Money Laundering, International Crime and Terrorism*. London: Kogan Page.

Morley, N. J., L. J. Ball, and T. C. Ormerod. 2006. “How the Detection of Insurance Fraud Succeeds and Fails.” *Psychology, Crime & Law* 12(2): 163-180.

Newman, M. 2008. “The Physics of Networks.” *Physics Today,* November 2008

O'Gara, John D. 2004. *Corporate Fraud: Case Studies in Detection and Prevention*. Hoboken, NJ: John Wiley & Sons, Inc.

Phua, C., et al. 2005. "A Comprehensive Survey of Data Mining-based Fraud Detection Research."

Porter, David.; 2005. "The Evolution of Fraud Intelligence." In *Managing Information Assurance in Financial Services*, ed. Rao, H. R., M. Gupta, and S. J. Upadhyaya, Hershey, PA: IGI Global.

PRWeb. 2009. "Alarming Increase in Insurance Fraud." Available http://www.prweb.com/releases/2009/05/prweb2399334.htm.

Reuter, Peter, and Edwin M. Truman. 2004. *Chasing Dirty Money: The Fight Against Money Laundering*. Washington, DC: Institute for International Economics.

Risk & Insurance. 2009, "Insurance Data Shows Jump in Fraudulent Claims Linked to Recession." Available http://www.riskandinsurance.com/story.jsp?storyId=212728307&query=National%20Insurance%20Crime%20Bureau.

Robinson, Philip. 2007. "The FSA's Perspective on Insurance Fraud." Available http://www.fsa.gov.uk/pages/Library/Communication/Speeches/2007/0926_pr.shtml

Rowe, R., et al. 2007. "Automated Social Hierarchy Detection through Email Network Analysis." *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*. San Jose, California.

Weiss, Sholom, et al. 2005. *Text Mining: Predictive Methods for Analyzing Unstructured Information*. New York: Springer Publishing Company.

CONTACT INFORMATION

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

Name:	Terisa Roberts
Enterprise:	SAS Institute Ltd
Address:	Wittington House Henley Road Marlow, BUCKS SL7 2EB United Kingdom
Work Phone:	+44 1628 490 787
E-mail:	Terisa.Roberts@eur.sas.com