THE USE OF CREDIT SCORECARD DESIGN, PREDICTIVE MODELLING AND TEXT MINING TO DETECT FRAUD IN THE INSURANCE INDUSTRY

TERISA ROBERTS

MSc (PUCHO)

A thesis submitted in fulfilment of the requirements for the degree

PHILOSOPHIAE DOCTOR

in

Operational Research

in the

SCHOOL OF INFORMATION TECHNOLOGY

at the

VAAL TRIANGLE CAMPUS

of the

North-West University

Vanderbijlpark

Promoter: Prof Philip D Pretorius

2011

ACKNOWLEDGEMENTS

I would like to thank my study leader, Phillip Pretorius for his creative thinking, direction and inputs throughout this study and Ayesha Bevan-Dye for language editing. I would also like to thank the employees at the insurance organisations, credit bureaus and other organisations in South Africa with whom I worked for their practical insights and the generous manner in which they gave of their time and shared their knowledge. Furthermore, I wish to thank my employers for their support and encouragement in my quest to further my knowledge. A special thank you to my parents and sister, who have provided much needed direct and indirect support, and to my wonderful husband for his continual support, love and encouragement. In closing, I thank my Provider, the Lord Almighty, Jesus Christ, for giving me the opportunity to conduct this study - I am truely blessed.

To Him be the glory,

Terisa Roberts

SUMMARY

The use of analytical techniques for fraud detection and the design of fraud detection systems have been topics of several research projects in the past and have seen varying degrees of success in their practical implementation. In particular, several authors regard the use of credit risk scorecards for fraud detection as a useful analytical detection tool. However, research on analytical fraud detection for the South African insurance industry is limited.

Furthermore, real world restrictions like the availability and quality of data elements, highly unbalanced datasets, interpretability challenges with complex analytical techniques and the evolving nature of insurance fraud contribute to the ongoing challenge of detecting fraud successfully. Insurance organisations face financial instability from a global recession, tighter regulatory requirements and consolidation of the industry, which implore the need for a practical and effective fraud strategy.

Given the volumes of structured and unstructured data available in data warehouses of insurance organisations, it would be sensible for an effective fraud strategy to take into account data-driven methods and incorporate analytical techniques into an overall fraud risk assessment system. Having said that, the complexity of the analytical techniques, coupled with the effort required to prepare the data to support it, should be carefully considered as some studies found that less complex algorithms produce equal or better results. Furthermore, an over reliance on analytical models can underestimate the underlying risk, as observed with credit risk at financial institutions during the financial crisis.

An attractive property of the structure of the probabilistic weights-of-evidence (WOE) formulation for risk scorecard construction is its ability to handle data issues like missing values, outliers and rare cases. It is also transparent and flexible in allowing the re-adjustment of the bins based on expert knowledge or other business considerations. The approach proposed in the study is to construct fraud risk scorecards at entity level that incorporate sets of intrinsic and relational risk factors to support a robust fraud risk assessment.

The study investigates the application of an integrated Suspicious Activity Assessment System (SAAS) empirically using real-world South African insurance data. The first case study uses a data sample of short-term insurance claims data and the second a data sample of life insurance claims data. Both case studies show promising results.

The contributions of the study are summarised as follows:

- The study identified several challenges with the use of an analytical approach to fraud detection within the context of the South African insurance industry.
- The study proposes the development of fraud risk scorecards based on WOE measures for diagnostic fraud detection, within the context of the South African insurance industry, and the consideration of alternative algorithms to determine split points.
- To improve the discriminatory performance of the fraud risk scorecards, the study evaluated the use of analytical techniques, such as text mining, to identify risk factors. In order to identify risk factors from large sets of data, the study suggests the careful consideration of both the *types of information* as well as the *types of statistical techniques* in a fraud detection system. The types of information refer to the categories of input data available for analysis, translated into risk factors, and the types of statistical techniques refer to the constraints and assumptions of the underlying statistical techniques.
- In addition, the study advocates the use of an entity-focused approach to fraud detection, given that fraudulent activity typically occurs at an entity or group of entities level.

Keywords: *Insurance fraud, claims fraud, fraud detection, scorecard design, weights of evidence*

OPSOMMING

Verskeie navorsingsprojekte handel oor die gebruik van analitiese tegnieke vir die opsporing van versekeringsbedrog en tog toon dit wisselende grade van sukses wanneer dit praktiese geïmplementeer word. In die besonder, 'n paar skrywers beskou die gebruik van kredietrisiko-telkaarte vir bedrogopsporing as 'n nuttige analitiese opsporingsinstrument. Navorsing aangaande analitiese bedrogopsporing vir die Suid-Afrikaanse versekeringsbedryf is egter beperk. Verder, beperkings soos die beskikbaarheid en kwaliteit van data-elemente; hoogs-ongebalanseerde datastelle; interpretasie-uitdagings met komplekse analitiese tegnieke en die veranderende aard van versekeringsbedrog dra by tot die voortdurende uitdaging van die suksesvolle opsporing van bedrog. Versekeringsorganisasies is ook onderworpe aan strenger regulasies en ander faktore wat die druk om aandeelhouers tevrede te stel, vermeerder. Dit alles dra by tot die behoefte aan 'n praktiese en doeltreffende bedrogstrategie.

Gegewe die volumes van gestruktureerde en ongestruktureerde data wat beskikbaar is in datapakhuise van versekeringsorganisasies, sal dit sinvol wees vir 'n effektiewe bedrogstrategie om data-gedrewe metodes en analitiese tegnieke in ag te neem. Die gebruik van die analitiese tegnieke moet versigtig oorweeg word. Noukeurige oorweging moet beide die kompleksiteit en die moeite wat dit verg om die data voor te berei in ag neem, aangesien ‘n paar studies bevind het dat minder komplekse algoritmes gelyke of beter resultate lewer. Verder kan te veel vertroue in analitiese tegnieke die onderliggende risiko onderskat, soos waargeneem met kredietrisiko op finansiële instellings tydens die finansiële krisis.

‘n Aantreklike aspek van die struktuur van die kans ‘weights of evidence’ (WOE) formulering vir telkaartkonstruksie is die vermoë om data probleme te hanteer, soos ontbrekende waardes, uitskieters en skaars waardes. Dit is ook ‘n deursigtige en buigsame tegniek, aangesien die gegroepeerde veranderlikes aangepas kan word, gebaseer op deskundige kennis of ander oorwegings. Die voorgestelde benadering in die studie is om bedrogrisiko-

telkaarte op entiteitsvlak op te stel deur stelle intrinsieke en relationele risikofaktore in te sluit.

Die studie ondersoek die toepassing van 'n geïntegreerde assesseringsstelsel van verdagte aktiwiteite, deur gebruik te maak van empiriese data. Die eerste gevallestudie gebruik ‘n steekproef van korttermynversekeringseise en die tweede gevallestudie 'n steekproef van lewensversekeringseise. Beide gevallestudies toon belowende resultate.

Die bydraes van die studie word soos volg opgesom:

- Die studie identifiseer verskeie uitdagings met die gebruik van 'n analitiese benadering tot die opsporing van bedrog binne die konteks van die Suid-Afrikaanse versekeringsbedryf.
- Die studie stel die ontwikkeling van bedrogrisiko-telkaarte, gebaseer op WOE vir diagnostiese opsporing, binne die konteks van die Suid-Afrikaanse versekeringsbedryf, en die oorweging van alternatiewe algoritmes om splitpunte te bepaal.
- Om die diskriminerende prestasie van die bedrogrisiko-telkaarte te verbeter, het die studie die gebruik van analitiese tegnieke, soos teksdata-ontginning geëvalueer. Ten einde risiko faktore van groot stelle data te identifiseer, dui die studie daarop dat beide die aard van die inligting, sowel as die aard van die statistiese tegnieke in 'n bedrogopsporingstelsel noukeurig oorweeg moet word. Die aard van die inligting verwys na die kategorieë van die data beskikbaar vir ontleding, vertaal in risiko faktore, en die tipes statistiese tegnieke verwys na die beperkinge en aannames van die onderliggende statistiese tegnieke.
- Laastens, beveel die studie die gebruik van 'n entiteit-gefokusde benadering tot bedrogopsporing aan, gegee dat bedrieglike aktiwiteite gewoonlik voorkom op entiteitvlak of by 'n groep van entiteite.

Sleutelwoorde: *Versekeringsbedrog, eise bedrog, bedrogopsporing, telkaart ontwerp, ‘weights of evidence’*

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION AND RESEARCH QUESTION

'A problem well defined is half solved' – John Dewey

1.1 INTRODUCTION

Globally, fraud is a major problem for insurance organisations, regulators and policyholders. Moreover, insurance fraud tends to increase during economic downturns (LRP Publications, 2009; Leckey, 2009). According to the Kroll Fraud Report (2009), financial services are the hardest hit by an increase in fraud during the recent recession. Statistical fraud detection is one of several defences available to insurance organisations in their fight against insurance fraud (Bolton & Hand, 2002a). Statistical fraud detection makes use of the volumes of data residing within the data warehouses of insurance organisations and statistical techniques in order to identify patterns indicative of suspicious behaviour.

In the context of the South African insurance industry, limited research on statistical fraud detection is available and very few insurance organisations in South Africa employ statistical techniques for fraud detection. However, the South African Insurance Crime Bureau (2008) started a data collaboration initiative in 2008, with the mission to reduce fraud and prevent perpetrators from targeting more than one insurance organisation. The data-sharing platform should provide further opportunity for data-driven fraud detection.

The current chapter provides background on the state of the global insurance industry and elaborates on the extent of the global insurance fraud problem. This is followed by a discussion of insurance fraud in the context of the South African insurance industry. In the section thereafter, the traditional and advanced analytical techniques used for insurance fraud detection, with examples from the literature, are summarised and discussed. Within the context of the South African insurance industry, several challenges with an analytical fraud detection approach are identified, defined and highlighted in the following section. In the following sections, the research question, goals

of the study and research methodology are provided. The chapter concludes with an outline of the remainder of the study.

1.1.1 Current state of the global insurance industry

In recent times, the global insurance industry has experienced tighter regulatory requirements, together with widespread consolidation brought about by acquisitions and mergers (Deloitte, 2011). In addition, the global recession and new technological developments affected the insurance industry. In general, the importance of operational risk, of which fraud forms part, is increasing due to growing computerisation, acquisitions and mergers, and globalisation (Doff, 2007:73). Operational risk arises from inadequate information systems, operational problems, breaches in internal controls and fraud or unforeseen catastrophes that potentially result in unexpected losses (Basel Committee on Banking Supervision, 2004:137).

The industry factors and their relation to insurance fraud are described in more detail below:

Increase in regulatory requirements

International regulatory directives, such as Sarbanes Oxley and Solvency II (Doff, 2007:4), are contributing to the standardisation of the insurance industry worldwide, and this increase in legislation affects insurance organisations in South Africa, as well as those in the international community. In the context of a developing economy such as South Africa, the increasing cost of compliance places additional financial pressure on organisations (De Koker, 2007; Nakajima, 2007). However, it warrants mention that these compliance measures have also benefited South Africa in that they aided in the reduction of fraud and encouraged more foreign investment (De Koker, 2007).

Acquisitions and mergers

Recently, the global insurance industry experienced widespread consolidation as insurance organisations aimed to grow globally, whilst reducing their operational costs by acquiring or merging with other organisations (Deloitte, 2011). However, these larger organisations seem more susceptible to fraud, which, hypothetically, is because more violations occur as organisations grow

and personal and structural controls decrease (Barnes & Webb, 2003; Finny & Lesieur, 1982).

Global recession

Given the current global economic turmoil and contracting economies, financial crime is on the rise. In September 2008, the sub-prime mortgage crisis emerged as a global economic crisis that shook the world's financial system (Abrahams & Zhang, 2009:1). In the face of the continued economic downward spiral, new data from the National Insurance Crime Bureau in the United States of America uncovered an increase in the number of questionable claims related to insurance fraud cases (LRP Publications, 2009). The press in the United Kingdom echoed the increase in fraudulent activity, where Allianz Insurance reported that fraudulent claims had doubled in the first three months of 2009 as organisations struggle to survive in the recessionary climate (PRWEB, 2009). The South African press also highlighted an increase in fraudulent insurance claims, attributed to the global recession (Stokes, 2010). The fear is that this is also affecting the work of counter-fraud specialists, who need to understand the changing nature and extent of the threat.

New technological advances

The development of new technologies, such as web-based technologies, mobile technologies and the expansion of social media, creates new channels of business for insurance organisations. On the negative side, it also creates new opportunities for perpetrators, if the necessary controls are not put in place.

As a player in the international community, the South African insurance industry is not immune to these hazards.

1.1.2 Definition of fraud

Before estimating the scope and magnitude of the global fraud problem, it is necessary to define what constitutes fraud. The legal definition varies by legal jurisdiction. In South Africa, fraud is defined under common law. The offence encompasses any unlawful act made with the intention to defraud, under which a misrepresentation is made that causes actual prejudice or which is

potentially prejudicial to another (Van Der Merwe & Du Plessis, 2004:483). The Fraud Act in the United Kingdom (Crown Prosecution Service, 2006) gives a statutory definition of the criminal offence of fraud, defining it according to three classes: fraud by false representation, fraud by failing to disclose information and fraud by abuse of position.

1.1.3 Sizing the global fraud problem

In the United States of America, the Coalition against Insurance Fraud (2011) estimates that insurance fraud costs at least $80 billion per year. The Comité Européen des Assurances (2007) estimates the cost of insurance fraud to be no less than 2 percent of the total annual premium income in Europe. In the United Kingdom, the cost of insurance fraud is estimated to be £1.9 billion per year (Insurance Fraud Bureau, 2011). In South Africa, the ratio of fraudulent claims to legitimate claims is higher than experienced in other countries, with fraud costing the short-term insurance industry an estimated R2 billion per annum. This is according to Samie, chairperson of the South African Insurance Association (Hartdegen, 2009). The official figure is that around 10 percent of insurance claims are fraudulent. However, Hartdegen (2009) quoted the Stakeholder Relations Manager of the South African Insurance Association, Pearson, as saying that these statistics are conservative and, in actuality, fraudulent claims amount to about 30 percent of the claims submitted annually. According to the Insurance Fraud Bureau (2011) in the United Kingdom, fraud adds 5 percent to the average insurance premium in the United Kingdom. In other countries, such as South Africa, it is estimated to add as much as 15 percent to the average premium.

However, these are only estimates given that it is virtually impossible to determine the exact value of the amount of money stolen through insurance fraud. By their very nature, insurance fraud crimes are designed to be undetectable, which means that a significant amount of fraudulent activity still goes unnoticed (Bolton & Hand, 2002a).

1.1.4 Context of fraud in the South African insurance industry

While South Africa has a comprehensive legal framework to combat financial crime, fraud remains under reported and there is a lack of resources devoted

to combating it. Personal safety is a serious concern in South Africa, so the question of priority, as raised by De Koker (2007), is that legal effort and energy tend to be utilised for more serious crimes than those committed by insurance fraud perpetrators. In a concerted effort to combat organised financial crime, the South African insurance industry collaborated in forming the South African Insurance Crime Bureau, with the support of the South African Insurance Association. The South African Insurance Crime Bureau (2008) was formed in 2008 to address organised fraud and crime in the short-term insurance industry, and to identify repeat fraudulent offenders that target more than one insurance organisation. Van Zyl (2010) of the South African Insurance Fraud Bureau states that specialist investigators, the police and other crime prevention agencies agree that a large proportion of insurance fraud is organised and perpetrated by syndicates. He quotes the Insurance Fraud Bureau in the United Kingdom who found that 40 percent of the insurance fraud believed to be opportunistic was also organised (Van Zyl, 2010).

Within the context of increasing regulatory demands, the globalisation of the industry and a global recession, it has become increasingly important for insurance organisations to adopt a comprehensive fraud risk-management strategy which moves away from a pure judgemental approach towards a data-driven approach.

1.2 USE OF AN ANALYTICAL APPROACH TO ADDRESS THE FRAUD PROBLEM

Advances in the use of analytical approaches for fraud detection include the use of statistics, visualisation, auditing, data mining techniques, artificial intelligence and others. Bolton and Hand (2002a) summarised the statistical techniques and challenges for fraud detection, not only for insurance fraud but also for fraud in general. In a rejoinder to this paper, Bolton, Hand, Provost and Breiman (2002) mention that multiple interconnected approaches seem more appropriate for fraud detection. Phua, Lee, Smith and Gayler (2005) summarised and categorised the available published data mining techniques for fraud detection. One criticism offered by Phua *et al.* (2005) is that the research focuses excessively on complex non-linear supervised approaches,

where, in the long run and using real-world data, less complex and faster algorithms produce equal if not better results.

In this section, several traditional and analytical fraud detection techniques are discussed with examples from the literature.

1.2.1 Traditional fraud detection approaches

Traditionally, insurance organisations used and still use business-generated alerts, such as whistleblower fraud hotlines, watch lists and diagnostic fraud indicators for fraud detection.

Whistleblower fraud hotlines

Most insurance organisations employ whistleblower fraud hotlines for tip-offs from the public and police as a first line of defence against fraud. These tip offs are typically investigated by the special investigation unit of the insurance organisation, using internal audit procedures.

Watch lists

Watch lists are used to match entities against available internal and external databases in order to identify individuals involved in organised crime. For example, members of the South African insurance industry collaborate to develop databases to assist in the detection of anomalous information at the claims stage (South African Insurance Crime Bureau, 2008). Morley, Ball and Ormerod (2006) state that these databases provide a way of verifying the information supplied by claimants and they allow organisations to assess whether claimants have a history of suspicious or similar claims. In addition, they provide repositories for sharing information about claims histories across organisations and with other parties. Morley *et al.* (2006) mention that some organisations restrict the use of such databases to specialist investigators, while other organisations have attempted to introduce their use into front-line claims handling.

Diagnostic fraud indicators

Diagnostic fraud indicators are red flag rules based upon industry and business experience, which describe factors believed to be indicative of potential fraud. These are used to identify suspicious behaviour when

information is collected from the claimant. The claims handler is typically responsible for completing a survey on the diagnostic fraud indicators before the claim is processed. As such, these indicators are usually based on the biased judgment of the claims handler.

A basic criminal profile includes the suspect's age, gender, place of residence, intelligence level, occupation, marital status, type and condition of the vehicle, motivating factors, and arrest record (Mena, 2003:20). Data pertaining to the demographics of the policyholders are usually available in the data warehouse. Some studies show that, for example, young male policyholders constitute a higher risk in terms of fraud (Subelj, Furlan & Bajec, 2010).

The problem with these red flag rules is that fraudsters learn the rules and loopholes very quickly. Moreover, the reliability of these flags is seldom tested against real data and acting on a single rule, without the context of the policy, customer or related entity view, complicates the work of the investigators. In addition, the completion of a set of fraud checks may inconvenience good customers. Commercial confidentiality prevents the publication of an exhaustive list of indicators. For illustration purposes, a short list of indicators from the literature is included in Addendum C. Data sources at insurance organisations, like the data warehouse and departmental data stores, may contain the required data to test for diagnostic fraud indicators. Traditionally, only binary indicators are used and when a large number of indicators are triggered, the suspicious activity is further investigated.

1.2.2 Analytical fraud detection approaches

Analytical fraud detection approaches are broadly grouped into two types: supervised and unsupervised. A supervised approach requires a supervised training dataset.

1. Unsupervised approaches

Unsupervised approaches are particularly attractive for fraud detection due to the rarity of the minority class, which makes supervised classification difficult (Bolton & Hand, 2002b). However, the specialist investigator is required to validate the performance of these models, as these measurements are not easily quantified. Unsupervised approaches include anomaly detection, text

mining, social network analysis and expert systems. Cluster analysis is another form of unsupervised learning. Cluster analysis may be used for segmentation, multivariate anomaly detection and profiling.

Anomaly detection

Anomaly detection refers to the detection of patterns in a given data set that do not conform to an established normal behaviour. Expert judgment or statistical techniques determine the thresholds. The detected patterns are called anomalies and are often translated into critical and actionable information. Anomalies are also referred to as outliers, abnormalities, deviations or exceptions. Anomalies are detected based on simple rules, such as a deviation from the expected distribution of the input, or more complex thresholds based on the dynamic profile of an entity. Cahill, Lambert, Pinheiro and Sun (2002) discuss some limitations with the use of thresholds. One limitation is that the thresholds need to be sensitive to several different factors in order not to set off too many false alarms. Profiling may be used to construct an outline of an entity's individual characteristics. For example, the profile of a new claimant is compared with the typical profile of a suspicious claimant. Fawcett and Provost (2002) propose profiling at account level, based on recent behaviour to detect credit card fraud. Juszczak and Adams (2008) suggest the use of unsupervised one-class classifiers at account level to detect credit card fraud. Brocket, Derrig, Golden, Levine and Alpert (2002) recommend the use of principal component analysis of rank-ordered attributes to detect insurance fraud. Other new technologies draw upon dynamic profiling approaches, based on the customer history (FICO, 2010).

Text mining

Robb (2004) estimates that 85 percent of corporate data is of the unstructured type. Text mining is the process that uses a set of algorithms to convert unstructured text into structured data objects. Techniques are available to deal with semantics, syntax, stemming, part of speech tagging and the identification of entities. In the insurance industry, unstructured textual data includes notes of accident descriptions by call centre operators and

subsequent examiners' interactions. Text mining may be regarded as an exploratory tool to discover meaningful information that resides in textual data fields like the claim narrative. Much of the text mining literature focuses on search engines and other information retrieval methods. In property and casualty insurance, literature on text mining is sparse (Francis, 2006). When applying text-mining methods, predictive models may experience substantial improvements in accuracy (Woodfield, 2004). This is achieved by integrating the quantitative results (frequencies and summary information of the qualitative concepts and synonyms) from the text-mining analysis to enhance the performance of the structured fraud risk models.

Social Network Analysis

In its simplest form, a network is a collection of points (nodes) connected by edges (ties). Nodes are the individual actors or entities within the networks and the ties are the relationships between these entities. Social network analysis is a study of relationships between entities, where the entities are individuals. Subelj *et al.* (2010) use social network analysis and an expert system to detect organised groups of fraudsters.

Expert systems

Expert systems are a type of artificially intelligent system, which store expertise concerning the subject matter in a knowledge base and attempt to solve problems in a manner that simulates the thought processes of a human expert. Major and Riedinger (2002) integrate expert knowledge and statistical information assessment for healthcare fraud.

Other advanced analytical techniques mentioned in the literature include agent-based modelling (Bonabeau, 2002). Bonabeau (2002) suggests that organisational simulation be used to identify operational weaknesses at agent or entity level where fraudulent activity may occur.

2. Supervised approaches

Supervised approaches for classification include any predictive modelling technique where a model is fitted on a sample of known fraud cases and legitimate cases. The system may also output reason codes that indicate relative contributions of various factors to a particular result. Viane, Derrig,

Baesens and Dedene (2002) compared the performance of several classification techniques on motor insurance claims fraud. These include logistic regression, C4.5-decision tree, k-nearest neighbour, Bayesian multilayer perceptron neural network, least squares support vector machine, naïve Bayes and tree-augmented naïve Bayes classification. They found that the differences in performance over the range of techniques are small. Linear logistic regression and the support vector machine performed very well, and the addition of more predictive inputs boost the performance of most models significantly. In other literature, techniques such as neural networks, are used where the relative importance of the inputs are measured (Viane, Dedene & Derrig, 2005). The interpretation of these models, such as neural networks, remains a challenge.

Bermudez, Perez, Ayuso, Gomez and Vazquez (2008) propose the use of a Bayesian dichotomous model with asymmetric link function to deal with unbalanced datasets. Yang and Hwang (2006) propose the use of data mining on frequent patterns of clinical instances and feature selection to detect healthcare fraud. The use of case based reasoning, which applies a suite of algorithms in combination for fraud diagnosis in the credit approval process, is proposed by Wheeler and Aitken (2000).

Some authors suggest the use of supervised meta-classifiers for fraud detection, where an ensemble of models is used to make predictions (Phua, Alahakoon & Lee, 2004). The final prediction is made by voting, however interpretation of the risk scores remain challenging.

Scorecard design

Scorecard design is a supervised approach popular in credit scoring. Scorecard methodologies are widely used to assess the risk of providing credit to a particular consumer. A credit scorecard is a table with a set of risk factors or characteristics, where each characteristic consists of a set of attributes, with points associated with each attribute. The points are summed and compared with a decision threshold to determine the credit status of new applicants (application scoring) or existing applicants (behavioural scoring). Its popularity is attributed to the fact that the knowledge to apply and use the

credit scorecards are separated from the technical statistical knowledge required to build them (McDonald, Sturgess, Smith, Hawkins & Huang, 2011).

Thomas, Edelman and Crook (2002:169) state that the use of credit scorecards for fraud detection is similar to other uses in many ways. Viane, Derrig and Dedene (2004a) propose the use of a scoring framework with the Adaboost algorithm to detect insurance fraud. In the study of Viane *et al.* (2004a), only binary indicators at claim level are used. Credit scorecards are easy to interpret and flexible as the risk factors and bins are interpretable and adjustable based on expert knowledge. It is also predictive (Viane *et al.*, 2002) as it uses logistic regression to combine the risk factors into a predictive scorecard.

3. Hybrid approaches

A combination of supervised and unsupervised approaches is also used. For example, Williams and Huang (1997) use unsupervised clustering algorithms to prepare the data for supervised classification with decision trees. Williams (1998) proposes a genetic algorithm that allows the rules to evolve over time. Major and Riedinger (2002) use a combination of an expert system and statistical analysis to detect healthcare fraud. Hilas and Mastorocostas (2008) use a multilayer perceptron neural network and clustering techniques to improve an expert system to detect telecommunications fraud. To detect asset misappropriation, Jans, Lybaert and Vanhoof (2007) use unsupervised mapping of the purchasing process to identify hot spots and then supervised classification.

1.3 CHALLENGES POSED BY THE USE OF AN ANALYTICAL APPROACH TO FRAUD DETECTION

Although theoretical research papers abound, the use of analytical techniques to detect fraudulent activity has seen varying degrees of success in its practical implementation. Technological advances are only effective if used in conjunction with the experience of people and complimentary processes. Within the context of the South African insurance industry, the use of an analytical approach to fraud detection poses several challenges. The challenges are listed and described below.

1.3.1 Fraud is an n-class problem

Fraudulent events within the insurance business vary from opportunistic fraud to organised crime gangs. Fraud detection is typically regarded as a binary classification problem (fraudulent versus eligible claim) when in actuality it is an n-class problem, as most fraudulent incidents are unique. The literature distinguishes broadly between two main types of fraud - internal and external fraud. Although internal fraud detection should form part of a comprehensive fraud detection strategy and is often linked to external fraud, this study focuses on external fraud, where the crime is committed by an entity outside the organisation. For example, motor claims fraud may be committed at different stages of events and by different entities: applicants for insurance (new customers), policyholders (existing customers), agents, third-party claimants, and professionals who provide services. Common fraudulent activities include "padding" (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 ordinary people who want to cover their premium expenses, to technicians who inflate the cost of services or charge for services that were not rendered, to organised criminals and professionals.

To simplify the matter, distinction is made between internal and external fraud. For external fraud, the following three types of fraud exist (Baldock, 1997; Gracey, Collins, Jones, Plumb, Tilley, Welfare & Williams, 2009):

- Opportunistic offences (“the opportunist”)
- Repeat offences (“the amateur”)
- Organised crimes (“the professional”)

Figure 1.1 depicts the types of fraud. Opportunistic fraud is normally not pre-meditated. For example, after a legitimate loss, the claimant might inflate the claim to receive more than he/she is entitled to. However, on the other side of the spectrum, there are groups of individuals who invent schemes to defraud insurance organisations. According to Porter (in Rau, Gupta & Upadhyaya, 2007), organised networks of criminals, colluding with employees on the inside of organisations, present the greatest threat, for example, groups of

criminals who stage accidents to claim for large and illegal damages. Sadly, more often than not, syndicates use the illegal gains from insurance fraud to fund other crimes.

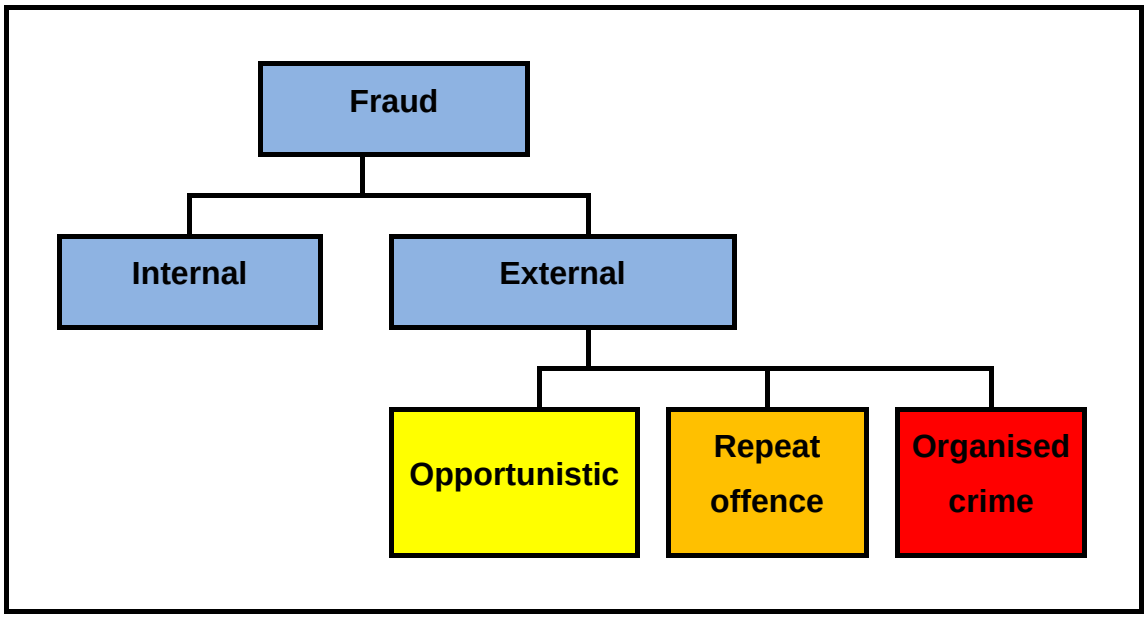


Figure 1.1: Hierarchical chart of the different types of insurance fraud

1.3.2 After-the-fact investigation of suspicious activities

Galimi and Earley (2005) emphasise the importance of using pro-active measures to combat fraud. They state that organisations must invest in predictive and preventive measures to mitigate losses. At present, fraud screening typically takes place initially when the claims information is provided to the claims handler.

Insurance organisations rely on claims managers to detect fraud, who often lack a comprehensive view of the entities involved in the claim. Morley *et al.* (2006) state that fraud detection methods are integrated with the claims handling business processes. Fraud detection methods should be applied iteratively throughout the claims processing cycle.

Traditionally, insurance organisations followed an inductive approach to fraud detection (Albrecht, Albrecht, Albrecht & Zimbelman, 2009). For example, when a red-flag rule or anomaly is triggered, an investigation is commenced. Specialist investigators would run further queries and analysis to determine the nature of the anomalies and types of fraud. In other words, the specialist

investigators do not start with a specific fraud in mind; rather, the investigation leads them and they learn from their research.

A pro-active approach, on the other hand, starts with the types of fraud that need to be identified and then associates risk factors with it. It then determines whether indicators exist to identify the types of fraud.

1.3.3 Tension between the need to maximise profits and the need to invest in anti-fraud measures

Jou and Hebenton (2007) state that in Taiwan, insurance organisations have little incentive to detect insurance fraud because of two factors: first, organisations want to minimise the number of reported fraudulent claims in order to protect their brand and, secondly, future higher premiums across the pool of clients absorbs the losses incurred due to fraud. The same may hold true in other countries. Morley *et al.* (2006) state that fraud is difficult to prove legally, and may incur further costs when police and legal teams are involved, without any financial outcome for the organisation beyond the recovered claim.

In contrast, insurers that are able to stop fraud and reduce losses can improve their bottom lines by offering lower premiums, creating a clear competitive differentiator.

It should be kept in mind that, typically, the more advanced the analytical approach to detect fraud, the more intensive the analytical resources, data and the amount of effort required. This linear relationship is illustrated in Figure 1.2 with examples of fraud detection approaches.

In general, the investment required to adopt a new fraud strategy must be lower than the cost of the actual fraud being committed, otherwise it is unlikely to be sanctioned.

Figure 1.2: Linear relation between the complexity of analytical techniques and its data requirements

1.3.4 Lack of resources in specialist fraud-investigation teams

Owing to the high volumes of incoming claims in modern insurance organisations, it is impossible for the specialist investigators to screen each incoming claim for suspicious behaviour. Schiller (2007) demonstrated that auditing becomes more effective and overcompensation is reduced when insurers are able to condition their audits on the information provided by detection systems. Furthermore, Schiller (2007) argues that once suspicious activity is identified, the audit cost of the investigation will be lower than before, as the detection system will not only provide a probability of fraud but will also generate relevant information to assist the investigation process. A semi-automated pro-active claims screening facility will assist the routing of claims, for example, through claims handling workflows. Claims that demonstrate low risk are settled quickly with the minimum transaction cost, whilst claims that demonstrate high risk are required to undergo an

investigation process, involving a resource-intensive verification process to determine the legitimacy of the claim. A semi-automatic classification approach should therefore be sensitive to the costs associated with the misclassification of the claims (Viane, Derrig & Dedene, 2004b).

1.3.5 Skewed and small marginal class distribution

Supervised classification approaches require a labelled data set. Labelled fraud data is rare. Few insurance organisations keep track of the history of fraudulent claims in their databases, as it has not been designed to collect the information. Typically, the specialist investigators maintain a log of accounts, which contains a list of fraudulent cases and cases under investigation. The marking of claims as being fraudulent is usually a manual task and is often subject to various sources of error. Most supervised classification approaches require balanced datasets to train.

1.3.6 Fraudulent behaviour is dynamic

Fraudsters adapt their behaviour to evade the current methods of detection. Often, the nature of the fraud changes as a response to the detection measure that has been put in place (Bolton & Hand, 2002a). Therefore, the fraud risk scorecards should be flexible and updated regularly to address the evolving nature of fraud. The establishment of an analytical assessment system is not the end goal of the project but it should rather provide a framework for ongoing detection. According to Hand (2007), the Pareto principle applies to fraud detection, where, in his experience, the first 50 percent of fraudulent activity is easy to stop, but the same amount of effort is required to stop the next 25 percent, the next 12.5 percent and so on. Following Hand's statement, organisations with no fraud risk-assessment system in place need not start with overly complex and resource intensive algorithms to detect fraud. A simple, easy to use, adaptable algorithm would suffice.

1.3.7 Value of detection as a function of time

The detection measure should identify the fraud as quickly as possible to minimise the loss amount and minimise the audit cost. For example, during the claims processing phase, more information may become available to

identify potential wrongdoing. However, in the analytical assessment system, priority should be given to data available at claim origination to detect suspicious activity as quickly as possible.

1.4 RESEARCH QUESTION

In the previous sections, the magnitude of insurance fraud, current techniques and challenges posed by an analytical approach to fraud detection were discussed. As insurance organisations in South Africa mostly use traditional judgemental approaches to fraud detection, this discussion gives rise to the following research question.

The research question is: given the volumes of structured and unstructured data available in insurance organisations' data stores, is it possible to aid the claims fraud investigation process with the use of scorecard design, predictive modelling and text mining, within the context of the South African insurance industry?

1.5 GOALS OF THE STUDY

The research question from the previous section is translated into the following goals of the study:

- Identify and define the challenges associated with an analytical fraud detection approach, applicable to the South African insurance industry.
- Create a fraud risk assessment system into which fraud risk scorecards would fit, to address specific challenges within the context of the South African insurance industry.
- Apply analytical techniques, such as text mining, to uncover risk factors to improve the discriminatory performance of the fraud risk scorecards.
- Evaluate the application of a fraud risk assessment system at entity level using South African insurance data.

1.6 RESEARCH METHODOLOGY

The following research methodology was followed in this study:

1.6.1 Analysis of literature

A critical survey of the available published literature was performed on the following topics:

- Statistical insurance fraud detection (summarised in Chapter 1).
- The context and challenges of insurance fraud detection in South Africa (summarised in Chapter 1).
- The use of analytical concepts like scorecard design, predictive modelling and text mining (technical details in Chapter 2 and 3).

1.6.2 Design of a Suspicious Activity Assessment System (SAAS)

A framework, called a Suspicious Activity Assessment System (SAAS), was designed to incorporate both quantitative (from the data) and qualitative (from the experts) risk factors into which multiple scorecards would fit. The study explained the technical concepts with examples and provided performance metrics to measure the discriminatory power of the fraud risk scorecards.

1.6.3 Identification of risk factors to improve the performance of a Suspicious Activity Assessment System (SAAS)

The study evaluated techniques such as anomaly detection, text mining, and social network analysis to identify risk factors which may improve the performance of the fraud risk scorecards.

1.6.4 Empirical Investigation

The study evaluated the performance of the SAAS, using both short term and life insurance data. The first step was to create analytical base tables, then the risk factors were identified and evaluated and finally the fraud risk scorecards were constructed and evaluated.

1.7 OUTLINE OF THE REMAINDER OF THE STUDY

In light of the research methodology, the remainder of the study is organised as follows. In Chapter 2, the study describes a proposed approach and design of a SAAS. It also provides details on the technical algorithms and performance metrics. Given the limited amount of research available on the topic of the identification of fraud risk factors, the study includes a chapter on

the topic in Chapter 3. Chapter 3 discusses the details of detecting anomalies, employing text mining algorithms and network information to identify risk factors. Chapter 4 provides an overview of the empirical investigations. Chapter 4 reports on the findings of two case studies which illustrates the performance of a SAAS using real-world insurance data. Chapter 5 concludes the study with a summary of the results, key findings, contributions, and outlines topics for further research.

CHAPTER 2

SUSPICIOUS ACTIVITY ASSESSMENT SYSTEM (SAAS)

'Plurality should not be assumed without necessity'
William of Ockam

2.1 INTRODUCTION

In the previous chapter, the scope and extent of insurance fraud and the challenges with an analytical approach to fraud detection were discussed. Keeping in mind the challenges, as described in Chapter 1, the current chapter puts forward an approach that incorporates analytical fraud risk assessments within an overall SAAS. The analytical fraud risk assessments are unable to operate independently from the business processes. As it is the case with a comprehensive credit risk-management strategy, an effective SAAS should be proactive, accurate, fast, flexible, consistent and transparent (Abrahams & Zhang, 2009:290). An over-reliance on black box analytical models can underestimate the underlying risk, as observed with credit risk at financial institutions during the financial crisis (Taleb, 2007:262). Transparency of the risk factors is imperative.

The system should leverage the volume of data that are available as well as the expertise and experience of the specialist investigators.

The first section provides a high-level overview of the approach. The approach provides a framework for analytical fraud detection. In the following section, the technical concepts of scorecard design using probabilistic weights of evidence (WOE) measures for risk factor identification and evaluation are described and motivated. The last section discusses useful validation measures to evaluate the performance of fraud risk scorecards for suspicious activity assessments.

2.2 PROPOSED APPROACH

A comprehensive fraud risk-management strategy should include measures to prevent, deter, recognise, detect and investigate fraudulent activity. The proposed approach supports a staged implementation approach, where one or several fraud risk scorecards, based on simple or complex risk factors, are

able to operate successfully. According to Mena (2003:296), a fraud detection methodology should not design a system that is fixed with hard coded rules and thresholds of which perpetrators may gain knowledge. The system should allow for the continual learning (or at least learning at regular time intervals) of the scorecards, regular monitoring and, if need be, revisions to adapt to the ever-changing characteristics of criminal behaviour.

As many and varied tools are required to address the fraud problem (Bolton *et al*., 2002), it is necessary to measure the effectiveness of a staged implementation approach, which broadens the usage of tools, techniques and risk factors with each iteration. Ideally, the assessment of entities should take place whenever a business event takes place (for example, at the point of acquisition, submission of a claim and at policy renewal).

Apparao, Singh, Rao and Bhavani (2009) provide a framework for financial statement fraud detection using a variety of data mining algorithms and stress the importance of feature selection. Adams (2010) recommends a multifaceted and multi-channel approach for fraud detection in general.

As outlined in Figure 2.1, this study has identified a six-step methodology for analytical fraud detection. The methodology shows similarities with widely used analytical model development methodologies (SAS Institute, 2009; Mena, 2003:296-299; Abrahams & Zhang, 2009:126).

The SAS data mining methodology; called the SEMMA methodology (SAS Institute, 2009); consists of five steps to cover the analytical model development life-cycle. The steps of the SEMMA methodology are Sample, Explore, Modify, Model and Assess. Mena (2003) proposes another data mining methodology, specific to fraud detection, called the CRISP-DM (Cross-Industry Standard Process-Data Mining). The methodology proposed by Abrahams and Zhang (2009:126) is specific to the assessment of credit risk, although its value to fraud detection resides in the fact that it puts a larger emphasis on the input from experts. The SAAS combines the characteristics of these methodologies into a single methodology applicable to analytical fraud detection within the context of the South African insurance industry. A key differentiator of the SAAS is that all six steps are heavily reliant on the

input and agreement from the subject matter experts. In addition, risk factor identification and scorecard construction are performed at entity level. Furthermore, the entity level views may be enhanced by the integration of external data sources or with the results from other predictive modelling techniques.

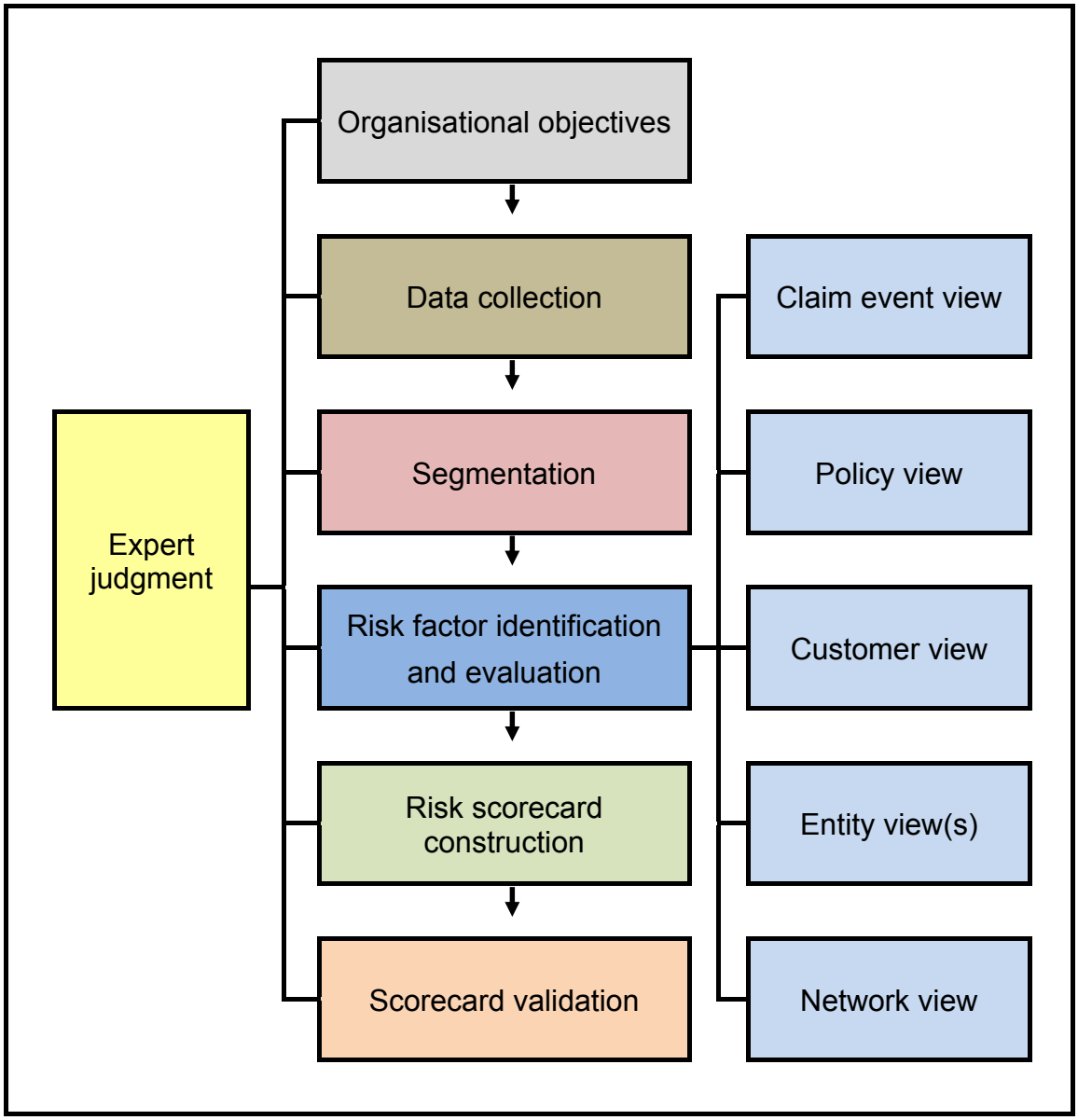


Figure 2.1: High-level process flow of a Suspicious Activity Assessment System (SAAS)

The six steps are described as follows:

2.2.1 Identify the organisational objectives

The identification of the organisational objectives is an important first step to identify the types of fraud to be detected. The organisational objectives about

the investigations are typically unique to each insurance organisation. The types of fraud need to be identified and then a decision needs to be taken as to which of these types of fraud need to be detected by the system. The costs associated with false positive and false negative predictions (*cf.* Cost matrix in Table 2.2) need to be determined. In addition, pre-emptive actions and enforcement that will take place need to be identified (Mena, 2003:294). For example, pre-emptive actions might involve the re-evaluation or repudiation of a claim, whilst enforcement might involve the legal department and police. During this phase, consideration should be given to the validation tests of the fraud risk scorecards (*cf.* Section 2.2.6) and the reporting requirements of the system. The costs and benefits of the project should be estimated during this phase.

2.2.2 Collect the data

Insurance organisations have large volumes of structured and unstructured data on the subject of incoming claims available, providing the opportunity to investigate the existence of statistical patterns to detect fraudulent behaviour. Data resides in multiple data warehouses and transactional databases. Consideration should be given to data quality issues and the impact of data quality on subsequent analysis, as well as how the data will be accessed and whether data from external sources, such as government agencies, will be combined (Mena, 2003:42).

As is the case with any statistical model development process, a good understanding of the underlying data and relationships between data elements is essential. Whilst numerous data exploration techniques exist, these are not discussed in detail in this study. Distribution analysis (graphs and summary tables) are useful. Further steps to transform the data encompass techniques to improve the data quality of real-world data, deal with data issues like missing values and outliers and, in general, improve the predictive performance of the data.

As mentioned in Chapter 1, one of the challenges with fraud detection is unbalanced datasets. Several techniques indicate that for unbalanced data sets, some form of over sampling of the minority class, combined with under

sampling of the majority class improves the discriminatory power of supervised learning approaches (Kubat, Holte & Matwin, 1998). Weiss (2002) discusses several approaches to deal with unbalanced datasets.

Chawla, Bowyer, Hall and Kegelmeyer (2002) propose a Synthetic Minority Over-sampling Technique (SMOTE), where the minority class is over sampled by creating artificial samples that are derived from values of the k-nearest neighbours along the line segment of the feature vector (original sample), rather than simple random sampling with replacement. Padmaja, Dhulipalla, Bapi and Krishna (2004) combine SMOTE with random under sampling of the majority class. They go further to suggest that the extreme outliers be removed before applying SMOTE.

Stefanowski and Wilk (2008) suggest local over sampling of the minority class combined with the removal of noisy observations from the majority class using k-nearest neighbour classification.

2.2.3 Segmentation

A large number of risk factors may affect and describe the occurrences of fraud, which may differ depending on diverse segmentation rules. For example, different types of fraud may apply to different market segments, policy types and entity profiles.

In addition, accurate segmentation rules for the known fraud cases are required to classify the types of fraudulent events correctly. For example, internal fraud cases would need to be excluded from the analysis of external fraud and vice versa.

When performing segmentation to support unique risk scorecards for sub-segments of the business, it should be considered that with segmentation comes added complexity in terms of more scorecards to maintain and processes to align. Segmentation is worthwhile if the improvement in the performance of the segmented scorecards outweighs the performance of the overall scorecard, both in terms of predictive power and maintenance costs.

2.2.4 Identification and evaluation of risk factors

Identification of risk factors

Viane *et al*. (2002) indicate that unexploited potential resides in the use of structural and systematic data mining for fraud detection, which goes beyond traditional red flag rules. A large number of risk factors may affect and describe the occurrences of fraud. Practically, it is not possible to include all relevant factors in the analysis, as the data may not be readily available or adequately prepared. Chapter 3 covers some of the techniques to identify risk factors in more detail. In Chapter 3, anomaly detection, text mining and social networks are discussed.

The risk factors used in the empirical investigations in Chapter 4 are grouped into the following five categories:

- Demographic information
- Static information
- Dynamic information
- Text data indicators
- Network indicators

The input from specialist investigators forms an integral part of the identification of the risk factors. Insurance organisations should maintain and update lists of risk factors as more information becomes available.

Evaluation of risk factors

In order to ensure a parsimonious scorecard, the suitability of the risk factors needs to be evaluated for inclusion in the subsequent fraud risk scorecards. For scorecard construction, the inputs need to be relevant and not redundant. According to Siddiqi (2006:78-79), in credit scoring, a common technique to measure relevancy, at a univariate level, is by using the WOE measures and corresponding Information Value (IV). The following section provides more detail on the WOE technique. A wide variety of discriminatory power measures is available for univariate variable selection, such as the Kolmogorof-Smirnoff statistic or the Gini coefficient.

To reduce multicollinearity, analysts use correlation analysis, like Pearson's correlation tests, and variable clustering techniques to evaluate the

redundancy of variables. One variable clustering technique uses principal component analysis to find groups of variables that are correlated with each other and uncorrelated with variables in other clusters (SAS Institute, 2010:3-26). For scorecard development, the analyst then selects a representative of each cluster. This technique may prove particularly useful in an environment where scorecards need to be updated regularly, as is the case in fraud detection. Another area where the variable clustering technique may prove itself useful is from a business perspective. For example, variable clustering may be useful where high risk exists that the end-users may gain knowledge of the significant scorecard inputs. To update the scorecards regularly, the analyst selects different representatives of a cluster to recalibrate the scorecards without a large outlay of analytical resources to develop new scorecards. Alternatively, several representatives of each cluster may be collected, whilst the end-users do not have knowledge of the actual factors used in the scorecards. Once the inputs have been evaluated in terms of their relevancy and redundancy, the risk scorecard can be fitted.

2.2.5 Construct the fraud risk scorecards

As mentioned in Chapter 1, a wide variety of supervised techniques is available. Logistic regression is part of a category of statistical models called Generalised Linear Models (GLM), which are discussed in more detail. A GLM relates the expected value of the target variable to a linear combination of the predictive inputs via a 'link function'.

$G(y) = \beta.X$

In addition to the linearity assumption implicit in this equation, GLM theory assumes that the dependent variable is distributed by the two-parameter family of distributions, known as the exponential family. The exponential family includes a range of distributions such as the normal, gamma, binomial, Poisson and many others (Yan, Guszcza, Flynn & Wu, 2009). The two parameters are known as the canonical and dispersion parameter.

In ordinary least squares regression, the analyst is required to verify that the assumption of normality (in particular, homoscedasticity or constant variance) and that linearity on the additive scale is satisfied. In the broader GLM

framework, the normality assumption is replaced by the weaker assumption that the distribution of Y is from the exponential family and linearity is replaced by linearity on the scale of the link function (Yan *et al.*, 2009). For a binomial distribution, the variance function is Y(1-Y) and the link function is the logit.

Regression methods are widely used to describe the relationship between a response variable and one or more explanatory or independent inputs. Logistic regression is used where the response variable is discrete (two or more possible values) and, more commonly, where the response variable is dichotomous (binary).

The logistic function, on which the logistic regression model is based, is popular because it provides estimates ranging between 0 and 1 and has an S-shaped curve to describe the combined effect of several risk factors on the dichotomous dependent variable (Kleinbaum, 1994:6). Logistic regression assumes that the difference between the natural logarithms of the class conditional data density functions is linear in the inputs or predictors.

There are two main differences between logistic and linear regression. First, the conditional mean of the dependent variable, when using logistic regression, is constrained between 0 and 1 (to achieve this, a logit transformation is used). The second difference is that the binomial and not the normal distribution describe the distribution of the errors. The unknown parameters βj are usually estimated by maximum likelihood using a method common to all GLMs. The interpretation of the parameter estimate, βj, is that of the additive effect on the log of the odds for a unit change in the jth explanatory input.

The model has an equivalent formulation:

$$p_i = \frac{1}{1+e^{-(\beta_0+\beta_1 x_{1,i}+\ldots+\beta_k x_{k,i})}}$$

The derivative of p_i with respect to $X = x_1, \ldots, x_k$ is computed from the general form:

$$y = \frac{1}{1+e^{-f(X)}}$$

where f(X) is a function of X.

For fraud detection, several authors argue that, ultimately, less complex algorithms seem to outperform their more complex and non-linear counterparts (Viane *et al.*, 2002; Phua *et al.*, 2005).

Another practical benefit of the logistic regression model is that its interpretability will aid the acceptance of a statistical decision support tool within the business. The additive effect of the inputs and the use of the WOE measures enable the use of a scorecard format (Siddiqi, 2006:113-118). The scorecard format appeals to a broad range of users who do not have statistical knowledge.

With the basic GLM framework in mind, the offset feature is briefly discussed. An offset is simply an additional model input, a, whose coefficient is constrained to one.

$$G(y) = \beta.X + a$$

To avoid omitted variable bias, analysts commonly perform a preliminary step to remove the effects of, for example, inputs not included in the model (Yan *et al.*, 2009).

A further practical benefit of using logistic regression is that the adjustment for over sampling can be done using the offset feature (SAS Institute, 2010:2-29).

Once the logistic regression is fitted, the posterior probabilities can be translated into an additive scorecard format using standard scorecard scaling formulae (Siddiqi, 2006:113-116). The scorecards can then be used to score new incoming claims for fraud detection.

Woodfield (2005) proposes the use of a two-stage model, which predicts the probability of fraud and the expected loss. The expected loss would enable the specialist investigators to investigate only those claims where the loss amount exceeds the cost of the investigation.

A basic approach to associate a loss amount with a fraudulent claim is to combine the claim amount and the audit cost. A predictive model may be used to estimate the potential loss associated with a fraudulent claim. For the continuous response variable, the loss amount can be estimated using a

lookup table or supervised classification model like linear regression or decision tree algorithm. The potential loss estimation model is based on the assumption that larger estimated losses should receive priority when investigated. Woodfield (2005) warns that when fraudsters realise that a maximum threshold exists, they will adapt their behaviour to stay under the radar.

Examples of hypothetical fraud risk scorecards that are used to estimate the probability of fraud and the expected loss are provided in Figure 2.2. The fraud risk scorecards contain a wide set of risk factors. A risk matrix is provided as an example of a way to combine the likelihood (probability of fraudulent activity) with the severity (the estimated loss). In the example in Figure 2.2 and based on business considerations, the specialist investigators may prioritise the cases in the red cells marked E (Extreme) for immediate attention. These cases are almost sure to happen with dire consequences, followed by cases in the orange cells, marked H (High) and yellow cells marked M (Medium). Cases in the green cells, marked L (Low) may be fast-tracked for quick processing as both the risk of fraud and the severity is low.

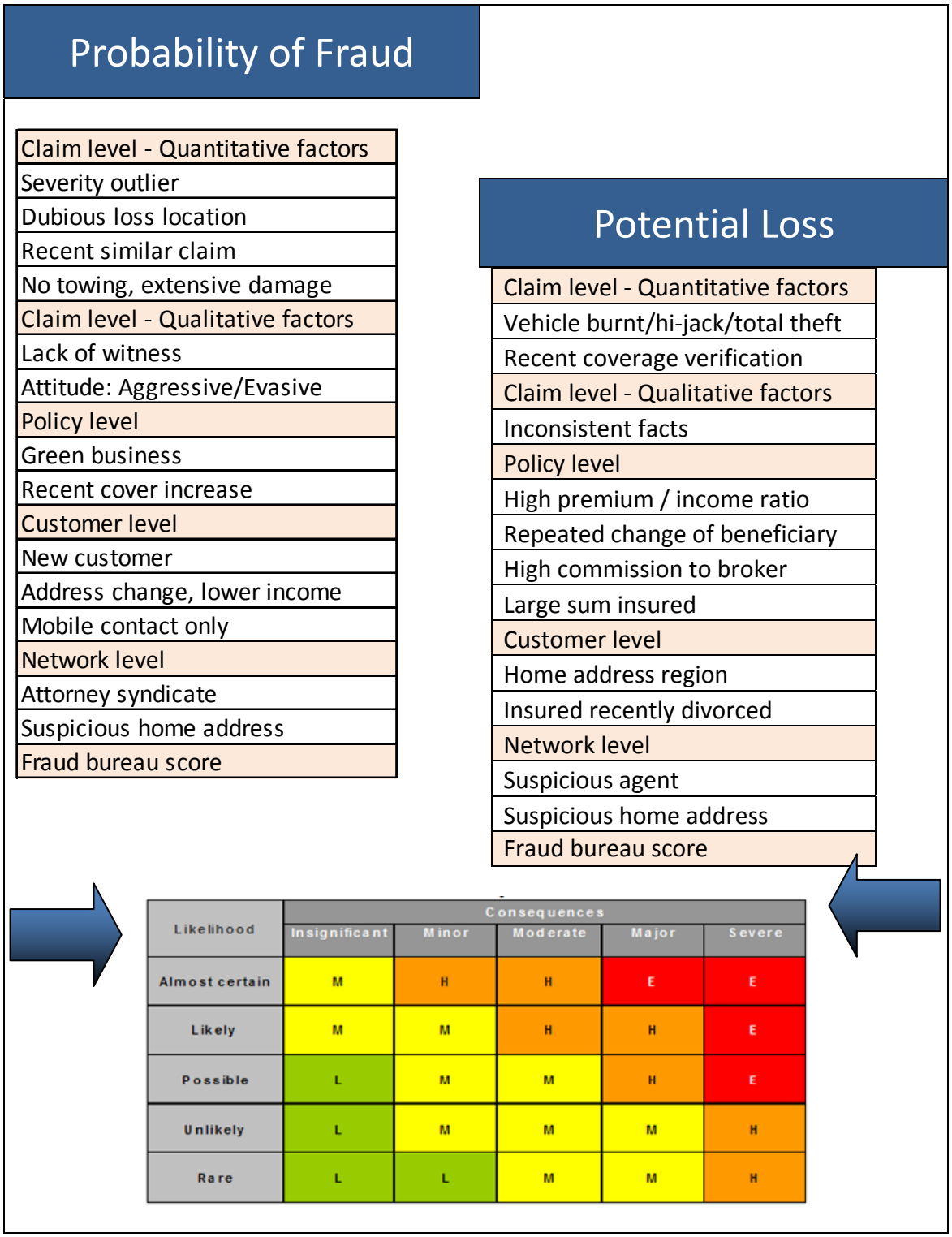


Figure 2.2: Hypothetical example of fraud risk scorecards at claimant level

One shortcoming of using supervised classification, as is the case with standard scorecard construction, is its lack of ability to identify novel fraudulent activities. The scorecards are only able to identify fraudulent activity as they have learned it from the past. Having said that, the use of fraud detection scorecards would allow organisations to make good use of the data they currently have, prevent repeat offences (one of the main types of external fraud) and close loopholes within their business processes that are abused by perpetrators.

As discussed in Chapter 4, the standard scorecards were enhanced with expert knowledge and the entities were evaluated based not only on their intrinsic characteristics but also on their relational characteristics; that is, the entities surrounding them and the networks of which they form part. These two properties make the fraud risk-assessment system slightly better equipped to identify novel fraud types. For example, the specialist investigators may learn from an external source that a criminal gang is targeting a specific region and the region is used as a significant input in the current scorecard. The analyst can then tweak the scorecard and associate manual WOE (i.e. very high risk) with the specific region. In addition, the network associations may identify entities that are regarded as low fraud risk, when evaluated empirically, but associated with other entities with high fraud risk.

2.2.6 Scorecard validation

Although the accuracy of the fraud risk scorecard is of interest, the main aim of a fraud risk scorecard is not to maximise accuracy, as the data is highly skewed. For example, if there are only 1 percent of known fraud cases in the data, a classification scorecard, which classes all cases as non-suspicious, will demonstrate a 99 percent accuracy rate, although it will not be very useful. Furthermore, the cost associated with a false positive and false negative is unequal and uncertain (Phua *et al.*, 2004). Scorecard accuracy measures go by different names, such as the percentage correctly classified (PCC), error rate, Receiver Operating Curve (ROC), and the like. The calibration of the scorecards needs to be evaluated as well, as the stability of the scorecards and the stability of the inputs may change over time.

In order to measure the discriminatory power of the fraud risk scorecards, the confusion matrix, the Receiver Operating Characteristic (ROC) curve and the Area under the ROC curve (AUROC) will be used. The cost/savings benefit of the scorecard may be evaluated using a cost matrix.

Note that to ensure that the scorecard is suitable for the population, the calibration of the scorecard need to be evaluated too. The Brier score and calibration plot can be used to evaluate the calibration of the scorecard. To measure the stability of the scorecard and input stability, the stability index can be used.

The importance of monitoring scorecard stability, input stability and population drift is critical for fraud detection. The use of a detection system will affect the behaviour of the fraudsters, called reactive population drift by Hand (2007).

2.2.6.1 Confusion matrix

The confusion matrix, as depicted in Table 2.1, contains information about the actual and predicted classifications (counts) done by a classification system.

All observations (N)	Predicted +	Predicted -
Actual +	True Positives (n_{TP})	False Negatives (n_{FN})
Actual -	False Positives (n_{FP})	True Negatives (n_{TN})

Table 2.1: Confusion matrix

The matrix can be used to measure the performance of the system and is used to derive the following performance statistics

$$\text{Accuracy Rate} = \frac{n_{TP} + n_{TN}}{N}$$

$$\text{Recall or True Positive Rate} = \frac{n_{TP}}{n_{TP} + n_{FN}}$$

$$\text{Precision} = \frac{n_{TP}}{n_{TP} + n_{FP}}$$

$$\text{True Negative Rate} = \frac{n_{TN}}{n_{TN} + n_{FP}}$$

If only the performance on the positive class (event class) is concerned, the F-measure was introduced by Lewis and Gale (1994).

$$\text{F-measure} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$$

When the performances of both classes are concerned, Kubat *et al.* (1998) suggested the use of the Geometric mean for unbalanced datasets.

$$\text{Geo-mean} = \sqrt{\text{Recall} * \text{Precision}}$$

2.2.6.2 Cumulative Accuracy Profile (CAP) and corresponding Accuracy Ratio (AR)

In order to construct the Cumulative Accuracy Profile, the observations are first rank ordered from riskiest to safest, based on the scoring rule probabilities (Engelmann, Hayden & Tasche, 2003:4). For a given fraction x of the total number of events, the percentage, d(x), of the events are calculated, which has a score equal or lower than the maximum score for fraction x. The CAP plots the cumulative distribution of the events (outcome variable equals one) against the cumulative distribution of the full sample, with the points connected with straight lines (interpolation). A perfect rating model would assign the lowest scores to the events, and would therefore increase linearly to one and then stay at one.

The Accuracy Ratio (AR) is defined as the area between the CAP of the scoring rule and the CAP of the random model (a_R), divided by the area between the CAP of the perfect model and the CAP of the random model (a_P). In other words, A_M is defined as:

$$A_M = a_R / a_P$$

Figure 2.3 depicts an example of a CAP.

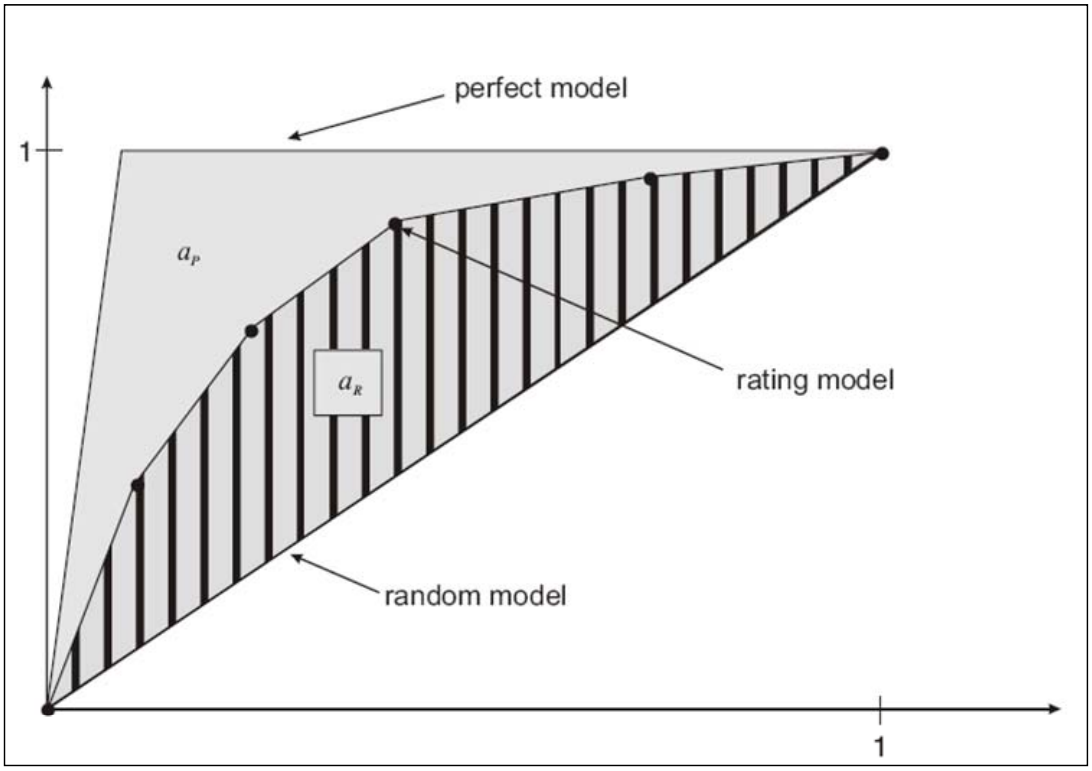


Figure 2.3: Example of the Cumulative Accuracy Profile (CAP)

2.2.6.3 Receiver Operating Characteristic (ROC) curve

The Receiver Operating Characteristic (ROC) curve plots the *1-specificity* (false alarm rate or 1 - true negative rate) on the horizontal axis versus *sensitivity* (hit rate or true positive rate) on the vertical axis for various values of a decision threshold (or cut-off values) imposed on the range of the scoring rule (Engelmann *et al.*, 2003:7). It evaluates the ranking ability of the scoring rule in terms of pairs of alternative operating conditions. Informally, the higher the ROC reaches to the upper left corner (point (0,1)), the better the ranking ability of the scoring rule.

Figure 2.4 depicts an example of a Receiver Operating Characteristic (ROC) curve.

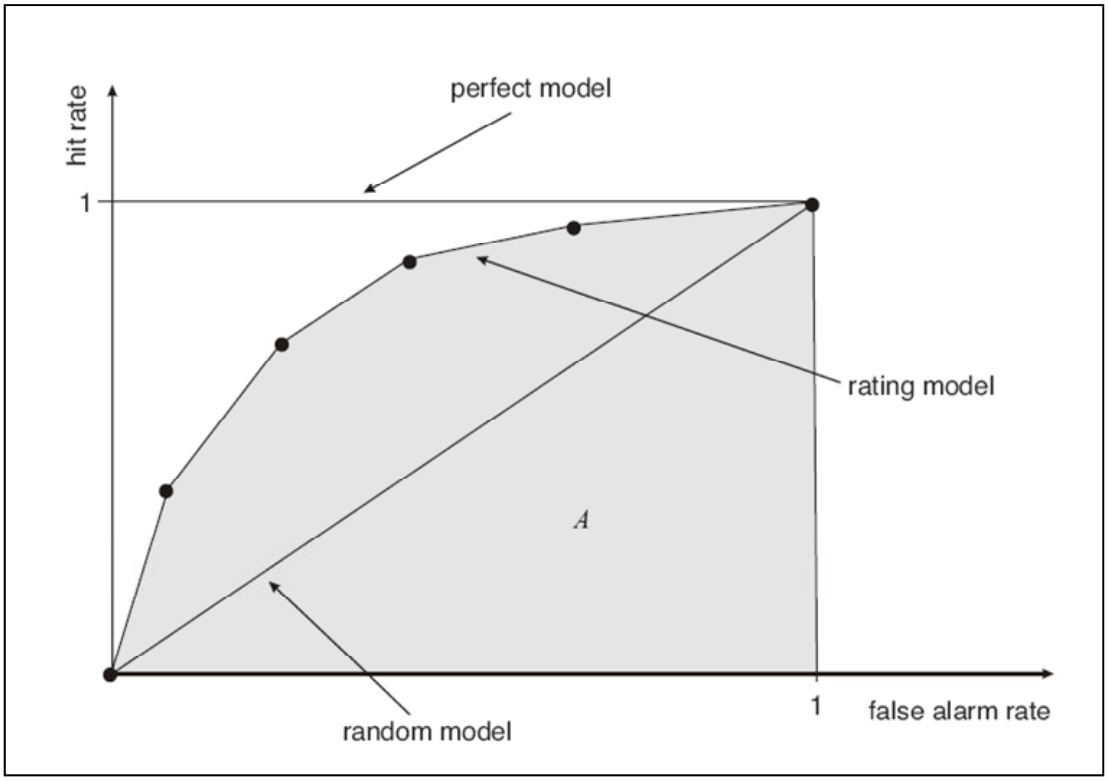


Figure 2.4: Example of a Receiver Operating Curve (ROC)

2.2.6.4 Area under the Receiver Operating Curve (AUROC)

The area under the ROC is a single-figure measure associated with the ROC curve. It is equivalent to the nonparametric Wilcoxon Mann-Whitney statistic, which estimates the probability that a randomly chosen positive instance is correctly ranked higher than a randomly chosen negative instance (Hand, 1997).

The AUROC and the AR are connected by means of a linear combination (Engelmann *et al.*, 2003:8):

AR = 2 * AUROC – 1

Consequently, neither AUROC nor AR depends on the proportion of events in the sample used for validation. According to Engelmann *et al.* (2003:6), the AUROC is 0.5 for a random model without discriminative power and 1.0 for a perfect model and between 0.5 and 1.0 for any reasonable rating model in practice.

2.2.6.5 Cost matrix

Chan and Stolfo (1998), and Fawcett and Provost (2002) argue that using measures like the error rate, percentage correctly classified or other ranking ability measure may not lead to optimal business decisions as the misclassification costs with fraud detection are unequal and the prior probabilities of the population are unbalanced. Some techniques for optimal decision-making, based on cost-sensitive classifiers, require well-calibrated posterior probabilities. It is therefore reasonable to evaluate the calibration of the fraud risk scorecard, which is the accurate estimation of the posterior probabilities, as well as cost-sensitive adjustments based on a cost matrix. The direct minimum expected cost classification associates an observation with the class with the minimum expected cost (Viane *et al.*, 2004b). An advantage of this technique, when using logistic regression, is that it does not require re-training of the predictive models, as costs are introduced after the learning of the models (Viane *et al.*, 2004b).

Table 2.2 depicts a typical cost matrix.

	Predicted +	**Predicted -**
Actual +	Cost $_{TP}$	Cost $_{FN}$
Actual -	Cost $_{FP}$	Cost $_{TN}$

Table 2.2: Typical cost matrix

A simplified version of the cost matrix is to suppose that accurate classifications incur no cost, whilst inaccurate classifications incur $Cost_{FN}$ and $Cost_{FP}$ (Hand & Vinciotti, 2003). The overall misclassification cost can then be denoted as:

$$OMC = n_{FP} * Cost_{FP} + n_{FN} * Cost_{FN}$$

where n_k is the number of observations classified as class k.

Note that with fraud detection, some auditing costs are incurred with true positives (Phua *et al.*, 2004; Chan & Stolfo, 1998), which are normally much smaller than the costs of false negatives, which are the costs of fraudulent

activity that go unnoticed. Viane, Van Gheel, Ayuso and Guillen (2004) propose the use of individual claim amounts and an estimated audit cost for each incoming claim to improve the profitability of claims screening.

One challenge to include costs is that the costs are often difficult to determine or to obtain. For example, some aspects of the cost of fraud to the organisation may be immeasurable in terms of monetary amounts, like previous undetected fraudulent activities, reputational damage, cross-departmental involvement and the like. Many organisations have only information on average costs, especially those who have not used a fraud detection system before.

It is therefore sensible to provide the user with a trade-off table to identify as many fraudulent events as possible, provided the false alarm rate is not too large, rather than a fixed cost matrix. The estimated values of the cost matrix can be used to calculate the cost or profit associated with each possible cut-off on the posterior probability range in the trade off table. The ROC curve (*cf* Figure 2.4) visualises the trade-off between hit rate and the false alarm rate using the trade-off table (Kubat *et al.*, 1998).

The importance of monitoring the scorecard calibration, stability, input stability and population drift is critical for fraud detection. Distance measures like the Brier score for scorecard calibration and stability index may be used to measure model stability.

As mentioned, the use of a detection system will affect the behaviour of the fraudsters, called reactive population drift by Hand (2007). Assuming that the new population of incoming claims are similar to the population of the development sample, the same predictive performance should be achieved as observed on the validation datasets.

2.3 WEIGHTS-OF-EVIDENCE (WOE) MEASURES

Univariate binning is a common method of removing variation while maintaining the relational structure in explanatory inputs. To reduce dimensionality and increase the predictive power of the input, both numeric and categorical inputs are binned.

It is thought that the WOE-based binning was first introduced by Good (1960). It is now a widely used technique in the area of credit scoring to support the construction of logistic-regression-based scorecards for retail credit scoring (Siddiqi, 2006:79-83).

2.3.1 Definition of weights-of-evidence (WOE)

The WOE measures are based on the log of odds calculation (Siddiqi, 2006:81).

A WOE measure is defined as follows:

Assume Y is a binary dependent variable (in the data it is coded as one for an event, fraud or positive instance, and coded as zero for a non-event or negative instance), with X an explanatory input variable with $x_{1, \ldots, x_k}$ discrete values of X.

The WOE measure for attribute i is calculated as follows:

$$WOE_i = \ln \left(\frac{|Y-_i|}{|Y-|} \Big/ \frac{|Y+_i|}{|Y+|} \right)$$

where $Y-_i$ depicts the number of non-event observations in attribute i and Y- depicts the number of non-event observations in the sample. Similarly, $Y+_i$ depicts the number of event observations in attribute i and Y+ depicts the number of event observations in the sample.

According to Siddiqi (2006:83), a more user-friendly way to denote the WOE for attribute i is:

$$WOE_i = \ln \frac{\text{Distribution of non-events}_i}{\text{Distribution of events}_i}$$

For variable selection, the IV or total strength of an input is calculated as the following sum of all attributes:

$$IV = \sum_{i=1}^{k} \left(\frac{|Y-_i|}{|Y-|} - \frac{|Y+_i|}{|Y+|} \right) * WOE_i$$

Siddiqi (2006:81) provides a rule-of-thumb interpretation of the IV as follows:

- < 0.02 unpredictive
- 0.02 – 0.1 weak
- 0.1 – 0.3 medium
- > 0.3 strong

As mentioned, using standard scorecard scaling formulae and the WOE (Siddiqi, 2006:113-116), the posterior probabilities from the logistic regression is translated into an additive scorecard format.

An example of WOE-based risk factors, their IVs and the corresponding fraud risk scorecard is provided in Figure 2.5.

Output Variables

Variable	Gini Statistic	Information Value	Level for Interactive	Exported Role
customer_riskfactor2	42.234	0.722	INTERVAL	Input
industry_class	36.806	0.61	ORDINAL	Input
customer_riskfactor1	41.315	0.53	INTERVAL	Input
claim_riskfactor1	32.192	0.449	INTERVAL	Input
claim_riskfactor2	30.585	0.341	INTERVAL	Input
network_riskfactor1	20.689	0.286	INTERVAL	Input
riskfactor9	22.171	0.237	ORDINAL	Input
pol_riskfactor1	26.054	0.225	INTERVAL	Input
riskfactor10	20.167	0.19	INTERVAL	Input
internal_rating	18.497	0.189	ORDINAL	Input
country_group	14.948	0.186	ORDINAL	Input
region_ind	14.468	0.09	NOMINAL	Rejected
customer_riskfactor10	10.042	0.082	ORDINAL	Rejected
riskfactor8	9.708	0.059	INTERVAL	Rejected

Scorecard

		Scorecard Points
claim_riskfactor1	claim_riskfactor1< -0.01	32
	-0.01<= claim_riskfactor1< 1.19	28
	1.19<= claim_riskfactor1	25
	Missing	13
claim_riskfactor2	claim_riskfactor2< 4.28	1
	4.28<= claim_riskfactor2< 4.58	16
	4.58<= claim_riskfactor2< 5.15, Missing	17
	5.15<= claim_riskfactor2< 5.98	61
	5.98<= claim_riskfactor2	27
country_group	1, 2	435
	3, 4	-33
	5, 7, 9, Missing	-36

Figure 2.5: Example of the Weights-of-evidence-based (WOE) risk factors, its Information Values (IVs) and corresponding fraud risk scorecard

2.3.2 Benefits of weights-of-evidence (WOE) binning

The benefits of binning (or grouping or bucketing) are of particular use to deal with real-world data challenges like missing values, outliers and rare classes. For example, missing values are typically binned into a separate class. The missingness of the input in itself may demonstrate a significant relationship with the outcome variable. Rare cases are binned together with bins with a similar fraud distribution. The binning process account for outliers too, in that the outlier is binned together with the high or low value bin and its impact is therefore considerably reduced.

Specifically, where the scorecards need to be updated regularly from a large number of available inputs, scorecard construction can be done with much less data preparation, whilst the results remain interpretable to support the investigation process.

To summarise, the benefits of the WOE-based binning of explanatory inputs are:

- It offers an uncomplicated way to deal with real-world data issues like missing values, outliers and rare cases.
- It explains the relationship between the explanatory input and the response variable, also where non-linear dependencies exist. The WOE does not necessarily need to increase monotonically with increasing values of the underlying input. Therefore, it permits the modelling of non-linear relationships between the input and the response variable.
- The analyst has control over the development process (Siddiqi, 2006:78)

For unbalanced datasets, the WOE measures improve the robustness of the scorecard as they use the class conditional probability measures.

2.3.3 Example of Weights of Evidence (WOE) calculation using credit data

In order to explain the WOE-based binning in more detail, it is necessary to look at a traditional credit-scoring example. Consider, for example, a numeric input: age of the applicant, used for credit scoring and obtained from a sample dataset with SAS software (Addendum B), which contains a balanced sample

of 3000 loan applications. The event or positive instance on the credit dataset is the default event. The histograms in Figure 2.6 depict the distribution of a numeric input: age, for the two values of the binary outcome variable: default (Y=1) and non-default (Y=0).

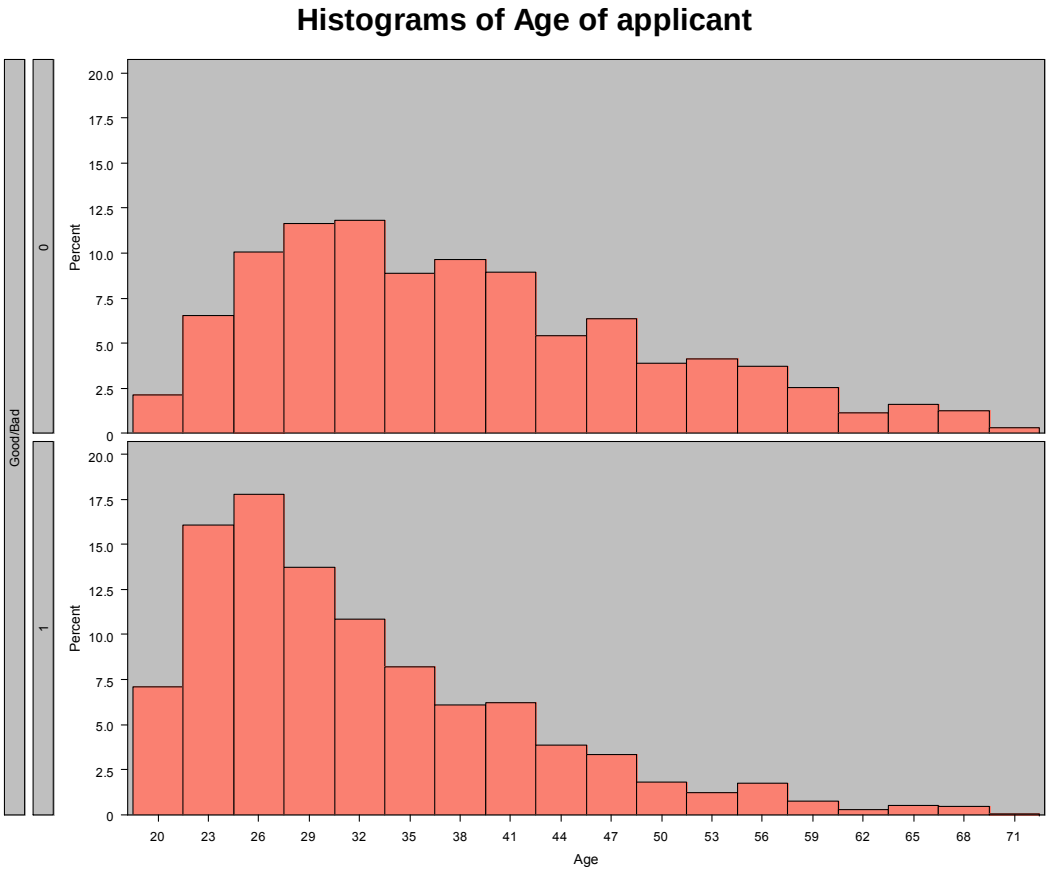


Figure 2.6: Histograms displaying the distributions of numeric input: Age of the applicant

The two qq-plots in Figure 2.7 confirm that the two distributions are from the same translation family (De la Rey, 2007).

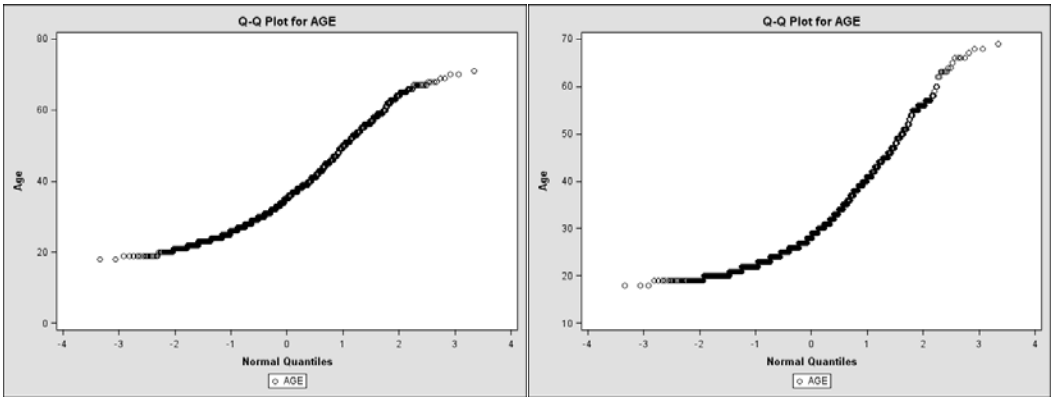


Figure 2.7: The qq plots for numeric input: age for good customers on the left and defaulted customers on the right

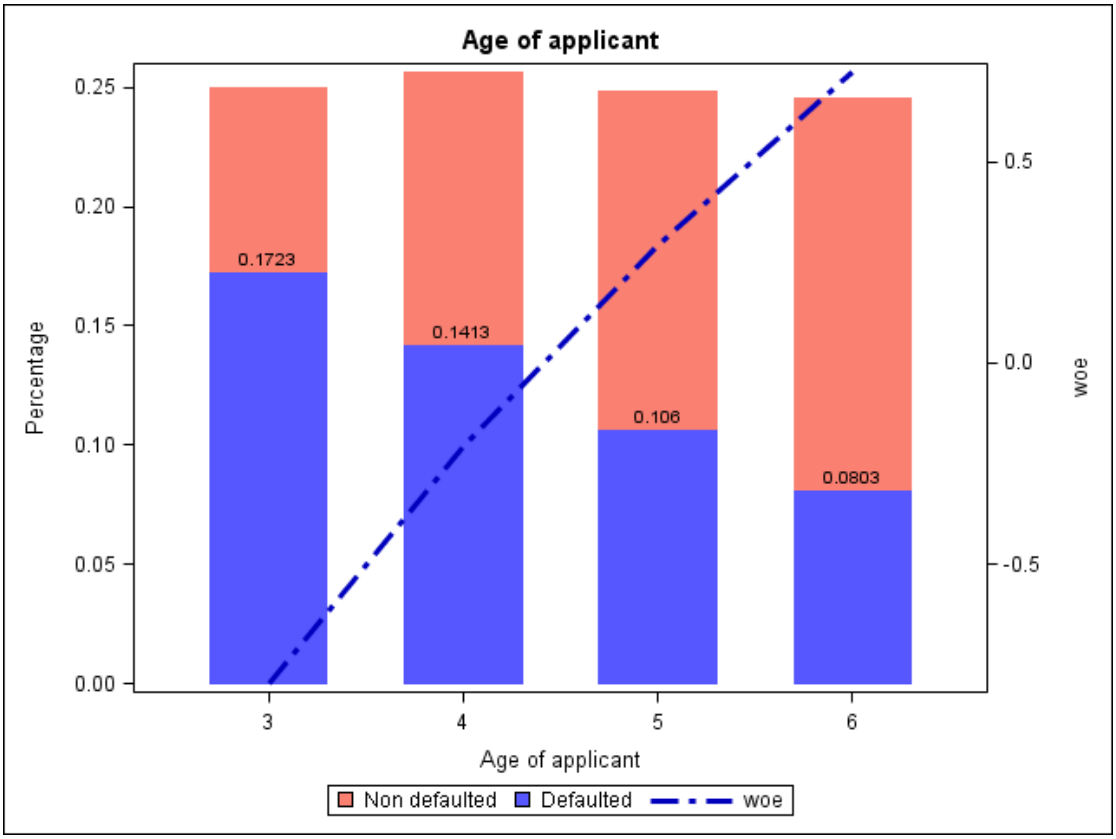


Figure 2.8: Weights-of-evidence (WOE) values of four bins of the age variable demonstrating decreasing credit risk across the range of age values

The graph in Figure 2.8 displays the WOE measures across the four quartiles of the age variable. Note that the split points were not optimised in this case. The large negative value of WOE observed in the first quartile corresponds to a large proportion of the event outcome (and vice versa, where high positive WOE measures correspond to low proportions of the event outcome).

Bin	Attribute	Total	Event	Event Distribution	Non-event Distribution	WOE
1	Age <= 25	750	517	0.345	0.155	-0.797
2	25 < Age <= 31	769	424	0.283	0.23	-0.206
3	31 < Age < 40	744	318	0.212	0.284	0.292
4	40 < Age	737	241	0.161	0.331	0.722
Information Value (IV)						0.306

Table 2.3: Frequency counts and weights-of-evidence (WOE) values of the four bins of the age variable

The non-uniformity of the WOE values indicates that a discriminatory relationship exists between the input and the outcome variable: the default event, which is confirmed by the IV of 0.306 (strong). The IV is calculated using the IV formula and the values in Table 2.3.

The linearity observed across the WOE values indicates that the risk of credit default decreases across the quartile range of age. In the example, a monotonic relationship exists between risk of default and the age of the applicant. In this example, the analyst used the WOE graph to determine the discriminatory power of the input and to evaluate the relationship between the input and the outcome.

In a similar way, the WOE values are calculated for categorical inputs.

2.3.4 Determining flexible split points

In the credit example in the previous section, the numeric input was discretised into quartiles or four uniform bins. Berzal, Cubero, Marin and Sanchez (2004) summarised several different techniques to obtain better bins for model estimation. It is common to use decision tree algorithms to determine better split points, in relation with the outcome variable, prior to the calculation of the WOE measures (Sarma, 2005; Siddiqi, 2006:78).

Traditional decision tree algorithms use the Entropy measure or Gini index to find better split points. However, as explained by Liu, Chawla, Cieslak and Chawla (2010), the Entropy measure, used to measure the information gain of a split point in decision tree algorithms, is biased towards the majority class distribution. Liu *et al.* (2010) shows that the bias applies to the Gini index too.

To find superior bins at a multivariate level, Hand and Adams (2000) suggested the use of simulated annealing or an exhaustive search across the range of all possible split points. Hand and Vinciotti (2003) motivate the importance of including the different costs associated with misclassification for unbalanced datasets. They put forward a strategy to include misclassification costs when estimating the parameters of the model. In order to explore the relationship of the inputs with the outcome variable, univariate analysis will be used.

Since fraud detection datasets are typically unbalanced, even after under sampling of the majority class, it is necessary to consider alternative binning algorithms at a univariate level. Cieslak and Chawla (2008) proposed the use of the Hellinger Distance (HD) to find better decision trees with unbalanced datasets. Liu *et al.* (2010) put forward another robust measure to deal with skewed datasets in decision tree algorithms, called the Class Confidence Proportion measure, which is similar to HD.

Liu *et al.* (2010) defines the well-known Entropy measure as follows:

$$\text{Entropy}_i = -\left(\frac{|Y-_i|}{n_i} * \ln \frac{|Y-_i|}{n_i} \right) - \left(\frac{|Y+_i|}{n_i} * \ln \frac{|Y+_i|}{n_i} \right)$$

where $Y-_i$ depicts the number of non-event observations in attribute i and n_i depicts the total number of observations in attribute i. Similarly, $Y+_i$ depicts the number of event observations in attribute i.

The Information Gain of a split is defined as:

$$\text{Information Gain}_{split} = \text{Entropy}_b - \sum_{i=1}^{k} \frac{n_i}{N} * \text{Entropy}_i$$

where N is the number of observations in the sample and Entropy_b is the baseline Entropy across all attributes.

The split that maximise Information Gain is selected.

Liu *et al.* (2010) define the positive Class Confidence (CC_+) of an attribute as:

$$CC+_i = \frac{|Y+_i|}{|Y+|}$$

where $Y+_i$ is the number of event observations in attribute i and Y+ is the number of event observations in the sample.

The CC_+ of an attribute corresponds to the denominator of the WOE measure.

According to Cieslak and Chawla (2008), as the HD is used to measure over a countable space rather than continuous, the HD can be rewritten as follows:

$$HD = d_H\,(Y+, Y-) = \sqrt{\sum_{i=1}^{k}\left(\sqrt{\frac{|Y+_i|}{|Y+|}} - \sqrt{\frac{|Y-_i|}{|Y-|}}\right)^2}$$

where $Y+_i$ depicts the number of event observations in attribute i and Y+ depicts the number of event observations in the sample. Similarly, $Y-_i$ depicts the number of non-event observations in attribute i and Y- depicts the number of non-event observations in the sample.

As it is not biased towards the majority class distribution, the study will evaluate the use of the HD hierarchical clustering to determine better split points across the range of attribute values for both categorical and numeric inputs. Note that some loss of information may occur as only univariate associations are considered. In the study, it is necessary to evaluate the risk factors at a univariate level to describe and understand their significance.

In practice, the split points of the bins are determined based on a combination of statistical analysis, business considerations and expert knowledge. Once bins have been identified using statistical analysis, the bins can be fine-tuned, based on business considerations (Siddiqi, 2006:79) and expert knowledge.

2.3.4.1 Example of the use of hierarchical cluster algorithms to bin categorical inputs

Greenacre (1993:41) proposes a data driven method to collapse levels of contingency tables based on the reduction of the Chi-squared test of association between the rows and outcome variable.

SAS Institute (2010:3-18) suggests that a clustering algorithm be used to collapse categorical inputs, with similar results to Greenacre's method.

In order to illustrate the binning of a categorical input based on Greenacre's method, an example will be used. Starting with a categorical input from the credit dataset with nine levels, the two levels with the least reduction in the Chi-square test are collapsed. The process is continued until all the levels are collapsed. The optimal number of levels is determined by selecting the number of levels that correspond to the minimum value of the log of the p-value of the Chi-square test. In this example, the optimum is three levels (as observed in Figure 2.9).

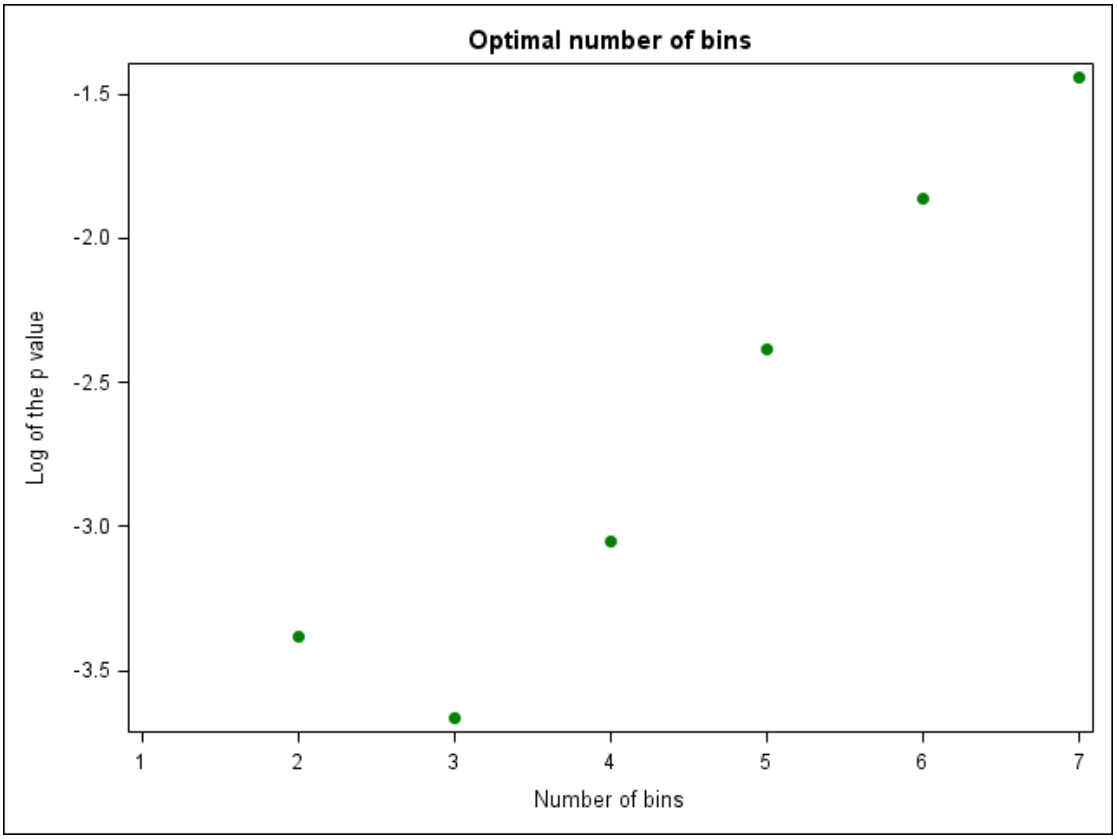
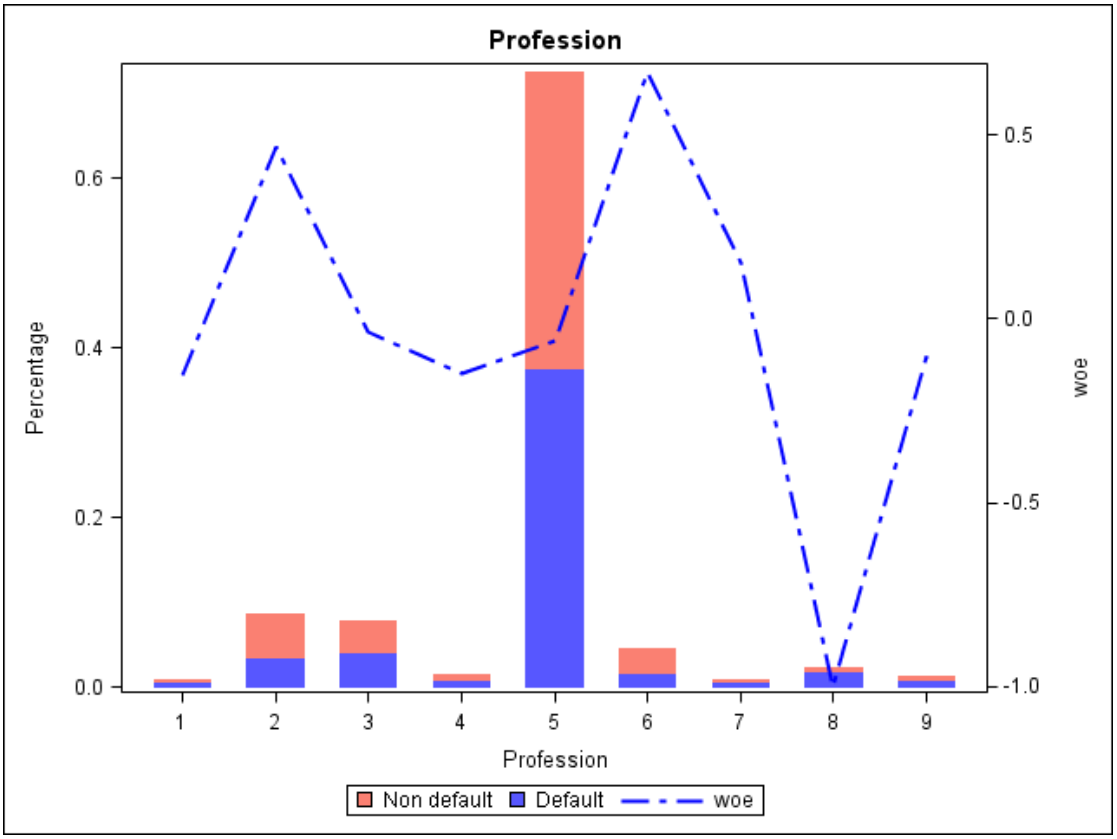


Figure 2.9: Log p-value across the range of number of levels



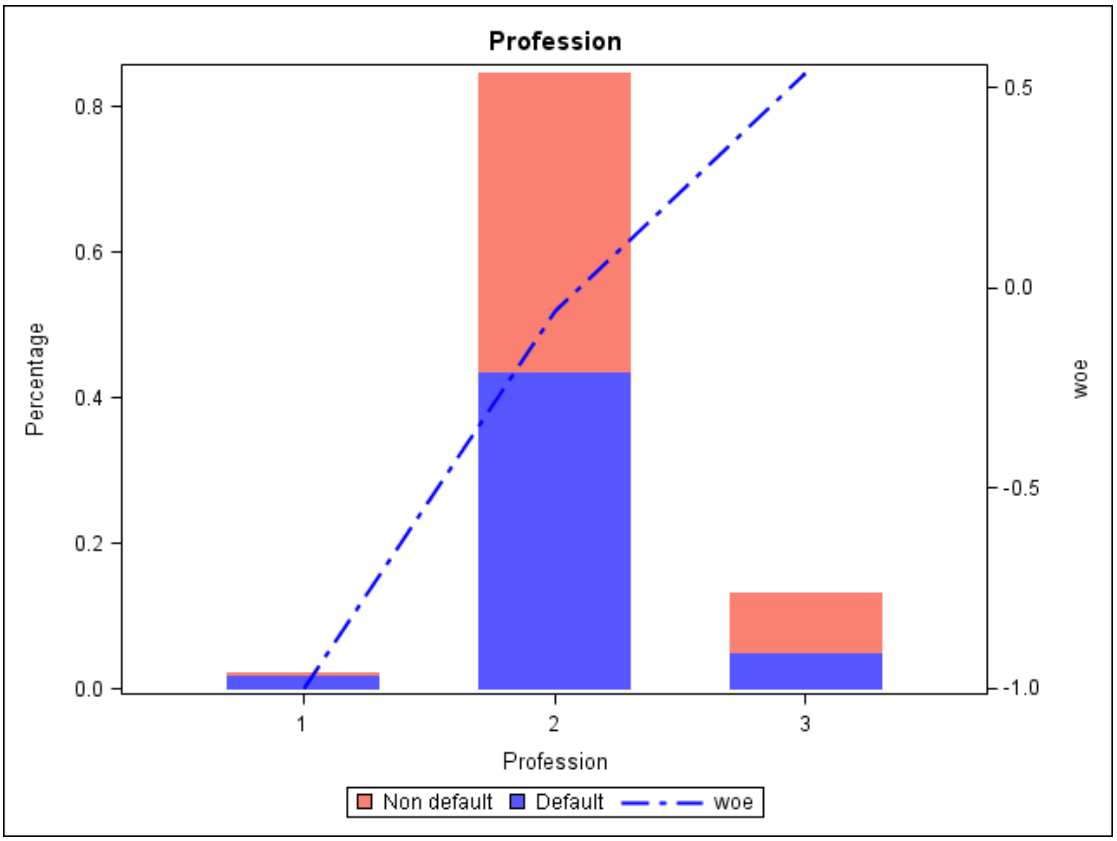


Figure 2.10: Weights-of-evidence (WOE) measures for the original and collapsed categorical input

Attribute	Total %	Event %	Event Distribution	Non-event Distribution	WOE
Profession Group 1	0.022	0.016	0.033	0.012	-1.00
Profession Group 2	0.846	0.435	0.870	0.823	-0.06
Profession Group 3	0.131	0.048	0.097	0.165	0.54
Information Value (IV)					0.060

Table 2.4: Frequency counts and weights-of-evidence (WOE) values of the three bins of the premium variable

The original IV of the nine-level input is 0.062 (weak). The IV for the three-level collapsed input is 0.060 (weak). It should be noted that, in this case, although the IV based on the Greenacre method is marginally weaker, it is still superior to the nine-level original input as the dimensionality is considerably reduced. Figure 2.10 shows the frequency distribution and WOE of the original input at the top and the collapsed input at the bottom. Table 2.4 display the WOE measures and IV of the collapsed categorical input.

For categorical inputs, an adapted version of the method proposed by Greenacre is proposed to reduce the levels of categorical inputs. The original method uses the event rate (ER) as cluster variable. However with unbalanced datasets, in some cases the Chi-square test may become unsuitable due to low frequencies in the levels. The Chi-square test may be replaced by Fisher's exact test (Liu *et al.*, 2010), although Fisher's exact test is computationally intensive in software packages. If the main aim of the scorecard is to find as many fraudulent events as possible, the cluster variable (and frequency term) may be replaced with the CC_+ (and event count). The clusters may also be collapsed using the HD as similarity measure, or using HD decision trees.

2.3.4.2 Example of the use of hierarchical cluster algorithms to bin numerical inputs

To find superior split points across the range of the numeric inputs, two aspects are considered: the number of split points and the positions of the split points. To compare the use of different similarity measures for numeric inputs, a hierarchical clustering algorithm adapted from the paper by Berzal *et al.* (2004) will be used. Berzal *et al*. (2004) collapse adjacent levels based on a similarity measure. The study will compare the use of the HD as similarity measure with the Entropy based decision tree.

In order to illustrate the binning of a numeric input in more detail, an example of a numeric input will be used. The explanatory numeric input is first binned into q quantiles, where the WOE, ER and HD are calculated for each bin. See the example in Figure 2.8 where four quantiles are used, where high negative values of the WOE (towards the left end of the scale) across the WOE range correspond with high fraud risk. However, the study wants to explore the use of the HD to find better split points in order to identify more of the minority class. The algorithm for a numeric input can be set out as follows:

1. Bin the input variable into q quantiles
2. Calculate the WOE, ER and HD for each quantile
3. Collapse the levels of adjacent bins based on a similarity measure, namely smallest difference in HD
4. Continue until a specific number of bins is reached
5. Evaluate bins and adjust based on expert knowledge
6. Calculate the WOE for each bin and use in subsequent analysis

The shortcomings of the use of the Entropy-based decision tree algorithm to determine split points with unbalanced datasets were discussed. The preliminary investigation evaluates the use of an alternative binning algorithm, namely the HD. In the investigation, the Entropy measure and HD were used to find better split points across the range of a numeric input.

For illustration purposes, an extension of the example in Section 2.2.3 is provided. The sample credit data set was used, which is a balanced dataset

with 1500 events and 1500 non-events. Observations from the class of events were randomly removed to create a new unbalanced dataset with 20 percent events and 80 percent non-events. For the elementary investigation, a single numeric input was selected to run the analysis. Starting off with the numeric input, age of the applicant, the input is binned into ten deciles (uniform bins). An Entropy-based decision tree is fitted with the number of branches set to a maximum of four. Based on the split points identified by the decision tree, the WOE measures are calculated and a logistic regression model fitted.

Similarly, starting off with the ten deciles, the HD measures are calculated for each decile. Adjacent levels with the smallest difference in HD are collapsed until four levels are left. Based on the split points identified by the hierarchical clustering algorithm, the WOE measures are calculated and compared with the results of the Entropy-based decision tree. Both the HD clustering and Entropy decision tree perform better than the WOE based on the four quartiles in terms of AUROC. The HD clustering and Entropy performed fairly similarly in terms of AUROC. The results are encouraging as the HD clustering algorithm is less complex than Entropy-based decision trees, however more testing is required. The results of the comparison are provided in Table 2.5.

Scorecard description	AR	AUROC
1. Quantile WOE	0.650	0.825
2. WOE (Entropy decision tree)	0.667	0.834
3. WOE (HD clustering)	0.662	0.831

Table 2.5: Comparison of the Weights of evidence (WOE) based logistic regression scorecards using a numeric input

Once a split is made on a binary input, no further processing is required in that dimension, prior to the calculation of the WOE measure. Informally, for a binary indicator, if the gradient of the WOE line is not equal to 0 and the IV is significant, the indicator may be useful for inclusion in further analysis.

2.4 CONCLUSION

This chapter put forward a methodology to implement a SAAS using scorecards at entity level and input from subject matter experts, like specialist investigators. The six steps of the methodology are as follows:

- In the first step the organisational objectives are identified.
- In the second step the data is collected from multiple source systems. A typical challenge within the fraud detection problem space is unbalanced datasets. The study mentions the use of SMOTE to improve the predictive performance of scorecards.
- In the third step the data is segmented based on the types of fraud, suspicious entities involved in fraudulent activities and other business considerations. The study mentions that segmentation may improve the predictive performance of the scorecards, although it may also add complexity and reduce the number of fraudulent cases available for scorecard construction.
- In the fourth step the risk factors are identified and evaluated. The study proposes the use of WOE measures and IV for this purpose. To evaluate the redundancy of inputs, correlation analysis or variable clustering techniques are proposed.
- In the fifth step, the scorecards are fitted. The study proposes the use of WOE measures and logistic regression to fit fraud risk scorecards.
- In the sixth and final step, the scorecard is validated by using discriminatory performance measures, such as the confusion matrix, CAP, ROC, AUROC and cost matrix.

In the next section, the study defined the WOE measures and IV and listed benefits of using WOE for scorecard construction. These include that WOE measures are robust against typical challenges like rare cases, outliers and missing values. The study also highlighted shortcomings from the literature, in the use of Entropy-based decision tree algorithms for binning inputs with unbalanced datasets. In situations where the learning algorithm needs to identify the more important minority class, an alternative binning algorithm,

like CC+, HD clustering or HD decision trees, may provide insight into the data otherwise undetected. The calculation of WOE measures were illustrated with credit scoring examples. In the example, both the HD clustering and Entropy decision tree (AUROC of 0.831 and 0.834 respectively) performed better than the WOE based on the four quartiles (AUROC of 0.825).

A limitation of the study is that the use of HD decision trees to identify superior split points with unbalanced datasets was not extensively tested. According to Cieslak *et al.* (2011), better split points should be obtainable using HD decision trees.

The next chapter will discuss the inclusion of potentially significant risk factors based on three types of analyses: anomaly detection, text data analysis and relational network analysis.

CHAPTER 3

IDENTIFICATION OF RISK FACTORS

'No man is an island' – John Donne

3.1 INTRODUCTION

Viane *et al*. (2002) commented that unexploited potential resides in the use of structural and systematic data mining for fraud detection, which goes beyond traditional red flag rules. Short-term and long-term insurance are dynamic and complex businesses. A large number of risk factors may affect and describe the occurrences of fraud. Some of these are generally applicable and published in the literature, whilst others may be unique to an individual organisation or niche industry sector. It is practically impossible to include all relevant variables in the analytical base table for fraud detection.

Furthermore, the data available for analysis is typically not usable in its raw form and often times incomplete, inaccurate and inconsistent (Berry & Linoff, 2000). Therefore, it is necessary to transform the available data into useful knowledge to support the fraud detection process. Anomaly detection as a canopy category for profiling, outlier detection and the application of Benford's law will be discussed in the first section. In the following section, the use of text data analysis is discussed to identify fraud risk factors. The last section introduces social network data analysis as a further area to be explored for risk factor identification.

3.2 ANOMALY DETECTION

Anomaly detection refers to detecting patterns that do not conform to an established normal behaviour. It is widely used for fraud detection. The thresholds are determined by expert judgment or statistical techniques. The patterns that are detected are called anomalies but are also referred to as outliers, abnormalities, deviations or exceptions. The definition of what constitutes normal behaviour is often times not precise. For univariate inputs, the anomalies are typically detected based on simple rules, like a deviation from the expected distribution of the input, or more complex thresholds based

on the dynamic profile of an entity. The following techniques are examples of anomaly detection.

3.2.1 Profiling

Industry experts, specialist investigators and psychologists have different opinions about the typical profile of a fraudster. According to Mena (2003:20), a criminal profile includes basic demographics like age, gender, place of residence, occupation and the like. Sutherland, Cressey and Luckenbill (1992:151-164) state that reported crime rates are consistently different for different age groups. Furthermore, they mention that gender strongly differentiate between criminals and non-criminals.

New technologies draw upon dynamic profiling approaches, which are based on the claims, policy and customer history (FICO, 2010). Artis, Ayuso and Guillen (1999) found that the historical behaviour of the insured is related to his/her propensity to commit fraud during the claims process. Weston, Hand, Adams, Whitrow and Juszczak (2008) use peer-group analysis to detect credit card fraud. They group similar accounts into peer-groups and detect anomalies using the Mahalanobis distance of the target account from its peer-group mean. For static profiling, the goal may be to summarise the data collected so far, and for dynamic profiling, the goal may be to predict the behaviour at the next observation. Extreme behaviour is often of more interest.

3.2.2 Outlier detection

As an illustration of outlier detection, it is useful to look at an example of an outlier trigger, called the *95_perc_outlier* of the claim amount by loss class. Figure 3.1 displays the claim amount (severity) distributions of several loss classes, which show the heterogeneity of risk between the different loss classes. Hossack, Pollard and Zehnwirth (1999:154) display similar results depicting the heterogeneity of risk within each of four risk categories. The trigger is set to flag claim amounts larger than the 95 percent percentile by loss class. Fraudulent activity is typically not detected based on a single trigger. However, when a combination of triggers is fired, the chance of

suspicious activity is all the more likely. Table 3.1 contains more examples of outlier triggers.

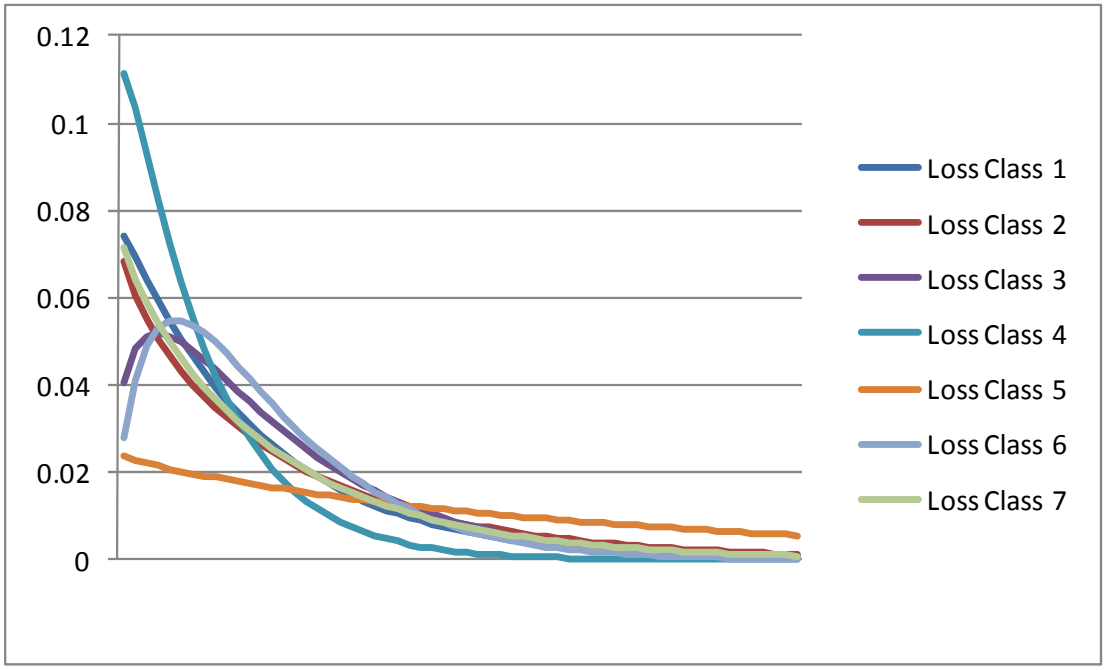


Figure 3.1: Heterogeneous Gamma claim amount distributions for several loss classes (special cut-offs are used for outlier detection)

Indicator	Description
loss_amt_outlier99	If the claim amount is larger than the 99% percentile by loss class
loss_amt_outlier95	If the claim amount is larger than the 95% percentile by loss class
claim_freq99	If the claim frequency is higher than the 99% percentile by group scheme indicator
claim_freq95	If the claim frequency is higher than the 99% percentile by group scheme indicator
tot_loss_policy99	If the total loss amount to date on the policy is larger than the 99% percentile by product
tot_loss_policy95	If the total loss amount to date of the policy is larger than the 95% percentile by product

Table 3.1: Binary anomaly indicators

3.2.3 Benford's Law

If the data of the population is distributed according to Benford's law, Benford's law can be used to detect anomalies. Nigrini (1999) recommends the use of Benford's law to detect tax declaration anomalies.

An astronomer, Newcomb (1881) was said to be the first to conclude that lower digits appear more often than higher digits. He derived formulae for the first and second significant digits as follows (zero digits are ignored):

$$p(d_1) = \log_{10}(1 + 1/d_1)$$

$$p(d_2) = \sum_{k=1}^{9} \log_{10}(1 + 1/(10k + d_2))$$

Benford (1938) found that the above digit distributions hold surprisingly well across various application domains. Hill (1995) derived the following formula for the joint distribution of the first and higher order significant digit:

$$p(D_1 = d_1, \ldots, D_k = d_k) = \log_{10}(1 + (\sum_{i=1}^{k} d_i 10^{k-j})^{-1})$$

where $d_i \in \{1, \ldots, 9\}$ and $d_j \in \{0, \ldots, 9\}$ and $j = 2, \ldots, k$

Benford's law is illustrated in Figure 3.2 where the claim amounts logged by two internal claims handlers with similar volumes of claims (the first under suspicion and the second not) are compared. The detected anomalies then become available for subsequent analysis.

Benford's law is more applicable to detect internal fraud (Debreceny & Gray, 2010). Another power law used for fraud detection is Zipf's law, which states that a pattern's frequency is inversely proportional to its rank. Huang, Yen, Luen-Wei, Yang and Hua (2008) uses Zipf's law for fraud detection by detecting anomalies in the frequencies of attributes. These attributes are not limited to quantitative inputs but may include textual strings and dates.

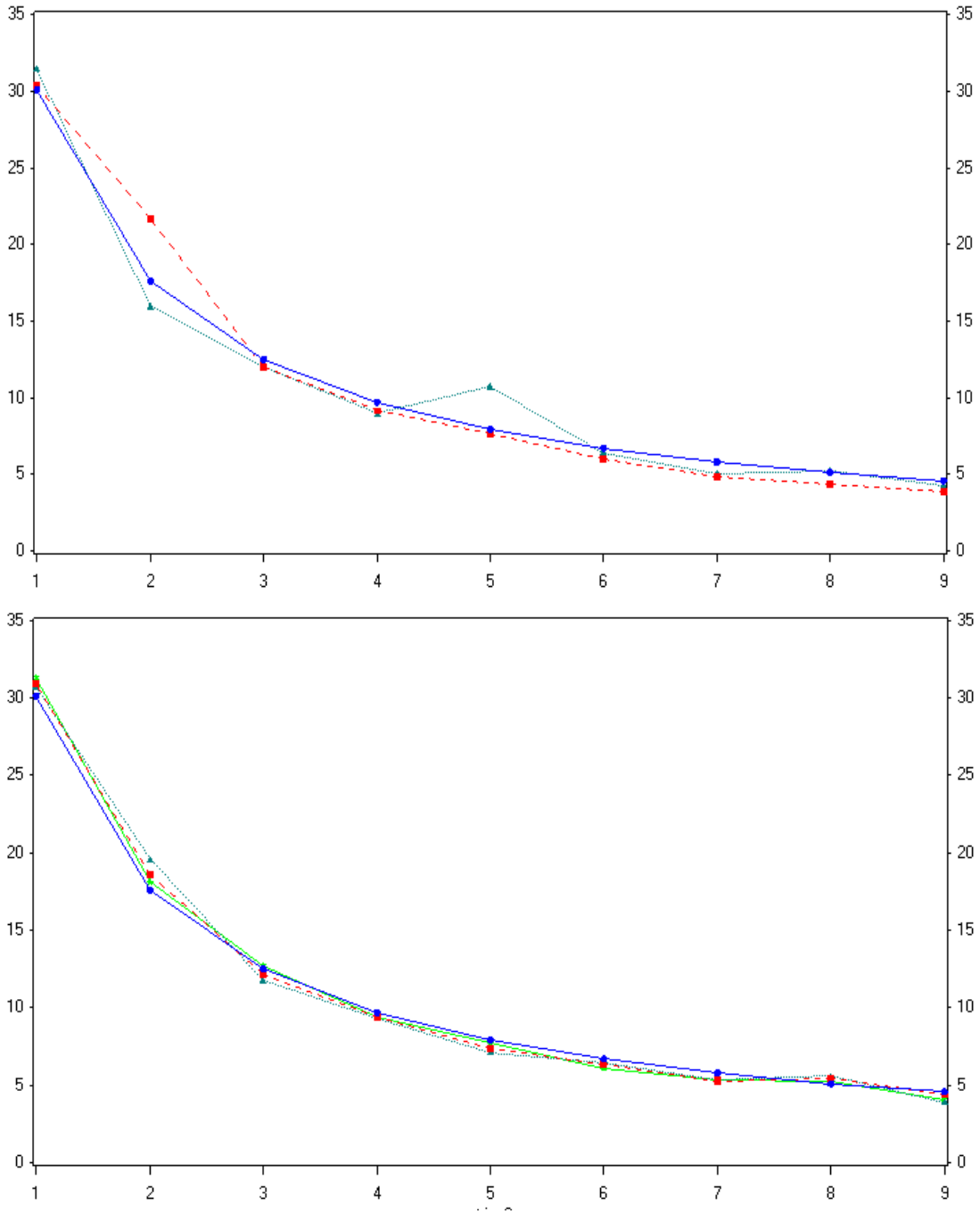


Figure 3.2: Distribution of the first digit of the claim amount of two suspicious employees (top) versus two non-suspicious employees (bottom)

3.3 TEXTUAL DATA ANALYSIS

3.3.1 Using text mining for predictive modelling

A large proportion of corporate data is of the unstructured type (Feldman & Sanger, 2006). The large size of insurance databases and the availability of textual fields (like the claim narrative for example) makes data mining on

textual data a useful option to identify risk factors pertaining to fraud. Text mining is the process that uses a set of algorithms to convert unstructured text into structured data objects. Techniques are available in most text mining software packages to deal with semantics, syntax, stemming, part of speech tagging and the identification of entities. Text mining may be seen as an exploratory tool to discover meaningful information that resides in textual-data fields, like the claim narrative.

Figure 3.3 depicts a graphical representation of the process to identify risk factors using text mining algorithms.

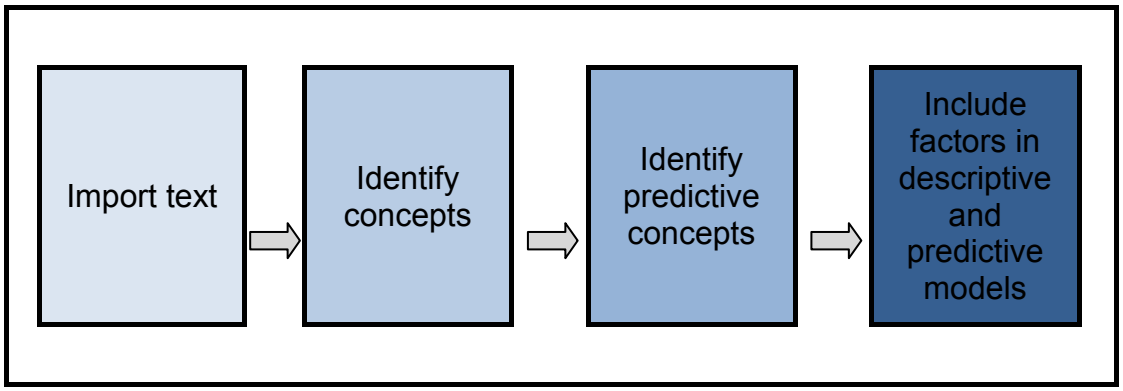


Figure 3.3: Process flow to identify risk factors in text data

One application of text-data analysis is to extract concepts from textual fields like the claim narrative. The textual field may contain information not available in the quantitative dataset. The first step is to import the free-form text fields into a text mining software package. Most text mining software packages have proprietary built-in algorithms to identify the main concepts. The important concepts are identified using synonym lists, proprietary algorithms and roll-up terms. Once the concepts are identified, the statistical significance of these need to be evaluated for inclusion in subsequent analysis. One approach is to compare the frequency distributions of the roll-up terms between the fraudulent claims sample and legitimate cases. These indicator variables called roll up terms, may be included in the subsequent multivariate analysis or further transformed to identify risk factors.

An example of a synonym list is included in Table 3.2.

Synonyms	Concept
hi jacked	hi-jack indicator
hi-jacked vehicle	hi-jack indicator
Burn	fire indicator
burnt out	fire indicator
Burned	fire indicator
W/off	write off
write off	write off

Table 3.2: Example of concept extraction using a synonyms list

3.3.2 Using text mining to improve data quality

As mentioned, real-world data are often times incomplete and inaccurate. The free-form text fields may contain important information, which might be missing or inaccurate in the structured data.

For example, in the first empirical case study in Chapter 4, text data algorithms, like the generalised edit distance, were used to match the names of the reported entities on the known fraud cases file to the data from the data warehouse. The generalised-edit distance is a generalisation of Levenshtein edit distance, which is a measure of dissimilarity between two strings (Navarro 2001). The Levenshtein edit distance is the number of deletions, insertions or replacements of single characters that are required to transform string one into string two.

3.4 SOCIAL NETWORKS

Fraud risk is present, owing to the fact that human relationships exist between an insurance organisation and its associates (called interested parties). Furthermore, professional and personal relationships exist between the interested parties themselves, which may not be evident at first observation. Lasalle (2007) states that due to the human factor; examining fraudulent activity is fundamentally different from auditing. The motives, rationalisation and opportunity to commit fraud differ, depending on which interested party or combination of interested parties are involved in the fraudulent activity

(Cressey, 1972). Interested parties have both intrinsic and relational properties, which might be indicative and descriptive of suspicious behaviour. The study identified the typical entities involved in an insurance claim, which is depicted in the diagram in Figure 3.4. The entities include policyholders, related parties (other policyholders sharing the same home address or home telephone number), agents, service providers, lawyers, assessors and other third party entities. In some instances, the value at risk may be much higher when fraud is committed at an agent or service provider level compared to policyholder level.

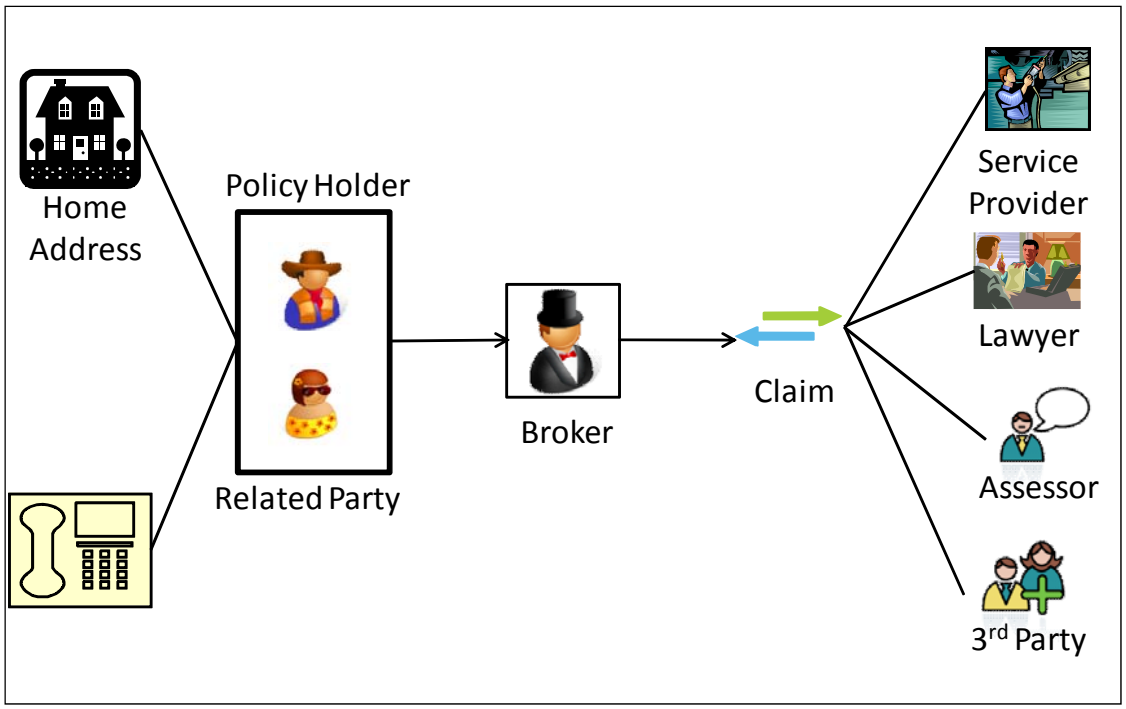


Figure 3.4: Entities typically involved in an insurance claim

Social network analysis is a study of social relationships in terms of vertices (nodes) and edges (ties). Nodes are the individual actors/entities within the networks and ties are the relationships between these actors. These relationships may be strong and obvious, such as a married couple sharing the same home address. The relationships may also display soft links, where entities demonstrate similar behaviour. Given the very large networks found in insurance data, it is necessary to detect communities or sub-networks. Hill, Provost and Volinsky (2007) build approximate social networks of very large networks and use collective inference to improve the predictive performance

of targeted marketing models. They mention that the technique would also apply to fraud detection. Data-driven techniques are available to collapse large networks into communities of networks (Blondel, Guillaume, Lambiotte & Lefebvre, 2008). Blondel *et al.* (2008) propose a technique for weighted networks, based on modularity optimisation. In a recently published paper, Subelj *et al.* (2010) use an unsupervised expert system and social network analysis to detect organised groups of fraudsters in three steps. First, construct and represent the networks of entities. Secondly, identify suspicious connected components (based on the structure of the sub-network). Thirdly, identify the suspicious entities in the suspicious connected components by iteratively assessing the entities based on their relational and intrinsic attributes until a threshold is reached. They found superior results by smoothing (Laplace smoothing) the suspicion scores based on the number of collisions an entity has been involved in. Their research focused on the identification of organised groups who stage accidents. This study proposes the use of a similar approach to construct networks but, in this case, the networks include policyholders and agents as the different types of entities and predictive modelling to generate fraud risk scores for the entities.

Figure 3.5 depicts a schematical representation of a proposed approach to construct networks.

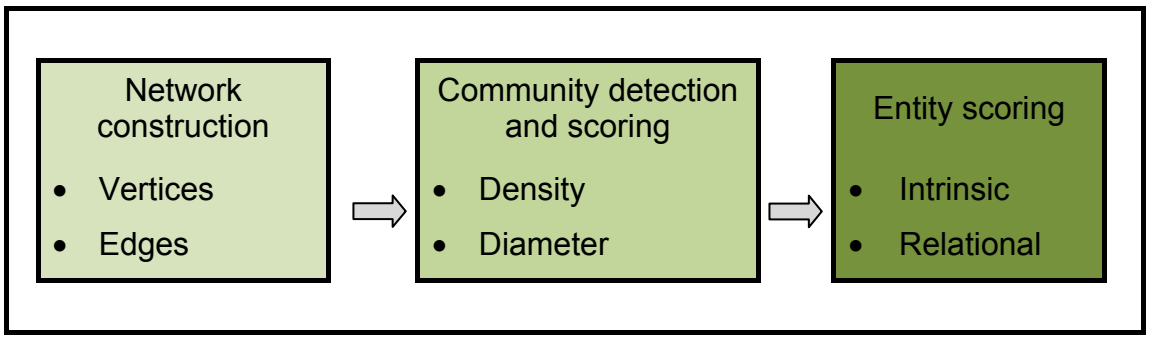


Figure 3.5: Steps to perform social network analysis with corresponding parameters

The first step is to construct the networks by identifying the vertices (or entities) and the edges (relationships between the entities). The next step is to identify communities or sub-networks in the large networks. Network metrics such as the density (amount of activity) and diameter (size) of the sub-

networks are calculated. The metric of density/diameter in itself can be used as a fraud risk factor to identify suspicious sub-networks. It is then necessary to score the entities in the sub-network and adjust their scores based on the network they form part of, as well as their first order and second order connections. Both the original scores and adjusted scores may be provided for further investigation. Figure 3.6 provides an example of network visualisation. Upon deeper investigation of the fraudulent links between entities, further suspicious activity and rings of criminals may be identified. The use of network visualisation is not a new concept in fraud detection (Porter in Rao *et al*., 2007); although limited research is available on the inclusion of the social network analysis results to improve predictive models like risk scorecards.

The human element of fraudulent activity should not be disregarded. An entity level detection system would support the work of specialist investigators. For example, say a traditional claims screening fraud detection system flagged a claim for suspicious activity accurately. It would be the work of the specialist investigators to determine whether the claimant or agent falsified the claim, whether the claims handler is committing internal fraud or whether the service providers are claiming for work not performed. The investigations of specialist investigators are lengthy and resource intensive. The delay in addressing the fraudulent activity might reduce recoveries. Even though the claim in itself were accurately classified as fraudulent, many false alarms were raised, which would inconvenience good policyholders, internal staff and other interested parties. The identification of the suspicious entity is important, but so is the context of the suspicious activity. Timely access to the behavioural history of suspicious entities and their networks would lead to shorter investigation periods, which in itself mean that potentially larger amounts may be recovered.

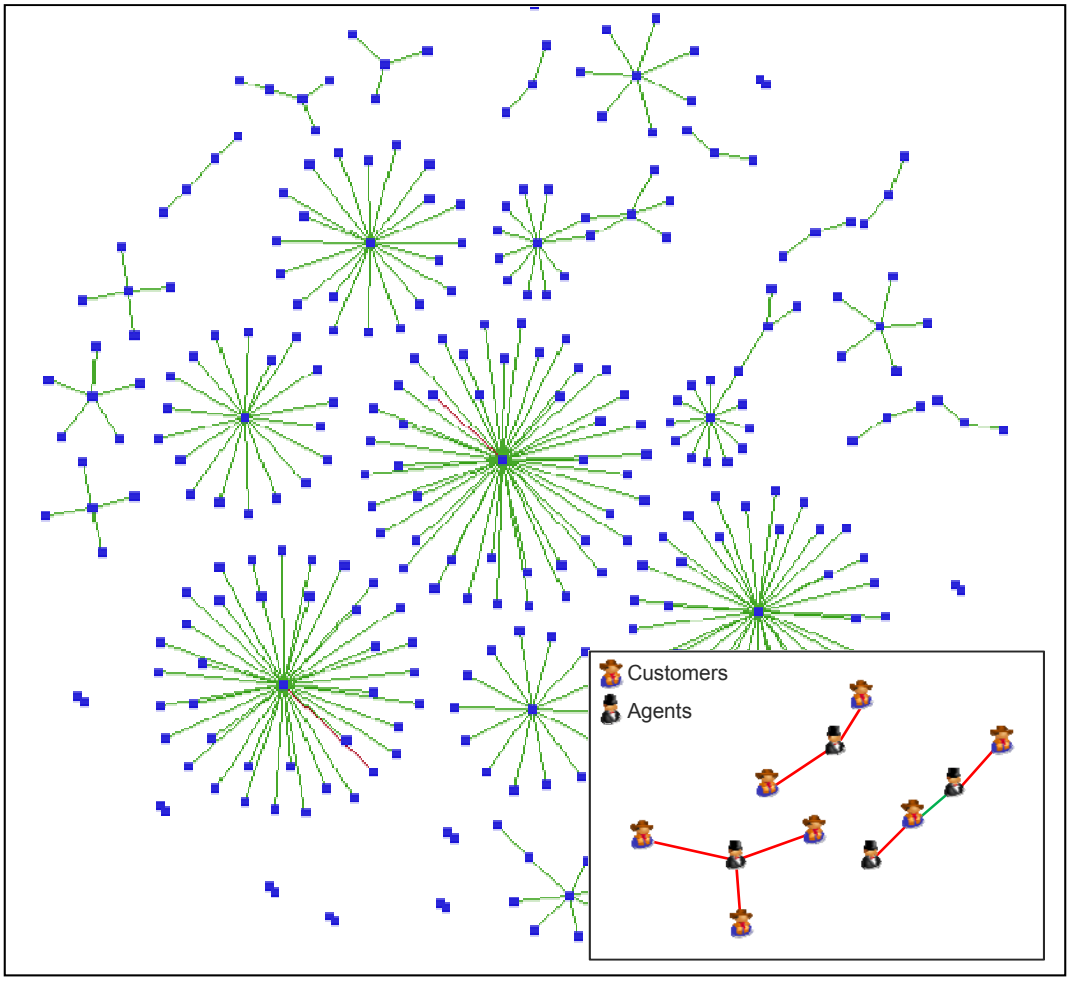


Figure 3.6: Network visualisation of a subset of agents and policyholders

Figure 3.6 provides an example of network visualisation. Upon deeper investigation of the fraudulent links between entities, further fraudulent activity and rings of criminals were identified: several agents are involved in more than one fraudulent activity, while some customers use more than one suspicious agent.

3.5 CONCLUSION

The current chapter provided details on three approaches to identify risk factors within the data stores of insurance organisations for fraud detection. These approaches include anomaly detection, text data analysis and social network analysis.

Anomaly detection, like outlier detection, is a widely used technique for fraud detection. Consideration should be given in order not to raise too many false

alarms. In addition to outlier detection, two additional data-driven types of anomaly detection methods were discussed, namely profiling and Benford's law.

The use of text mining algorithms as a data driven method to identify more risk factors for fraud detection were discussed. Text mining is a useful exploratory tool to summarise the information in free-form text fields, to identify key concepts and to uncover risk factors that would otherwise go unnoticed. It may also be used to improve the data quality of raw data, like a fuzzy merge on customer names of two separate source systems.

The last data-driven technique discussed in this chapter is the use of social network analysis and relational risk drivers to improve the performance of the fraud risk scorecards.

The following chapter will describe two case studies on actual insurance data where these approaches were utilised to detect insurance fraud.

CHAPTER 4

EMPIRICAL INVESTIGATION

'Intuition is a combination of empirical data observation,
and an ability to cut through the thickness of surface reality.'
Abella Arthur

4.1 INTRODUCTION

Bolton and Hand (2002a) mention that real-world data on insurance fraud is scarce. This scarcity may be attributed to the nature of the problem in that if organisations share their fraud strategies openly, perpetrators will also have knowledge of the rules. In addition, this scarcity may be the result of the reputational risk that exists when organisations disclose the extent of fraud within their organisations.

In the first section, as a preliminary investigation, an evaluation of the application of SMOTE on the credit dataset is provided.

In the following section, the study reports on two case studies that use the application of a SAAS on South African insurance data. The study provides examples of the calculations, although the publication of all of the results and details on specific risk factors are restricted by confidentiality.

The first case study uses short-term insurance claims (motor insurance) and the second case study life insurance claims (funeral products).

4.2 PRELIMINARY INVESTIGATION: EVALUATION OF THE SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE (SMOTE)

The credit data sample was used to evaluate SMOTE (Chawla, Bowyer, Hall & Kegelmeyer, 2002). The credit data sample is a balanced dataset with 1500 events and 1500 non-events. Observations from the class of events were randomly removed to create a new unbalanced dataset with 20 percent events and 80 percent non-events. Together with the binary event indicator, seven inputs from the dataset were selected and correlations evaluated. No strong correlations were observed. SMOTE was applied, to create 100 percent more fraud cases, although the new cases were synthetic.

First, a scorecard was developed using the unSMOTEd dataset. For scorecard construction, each input was discretised into five uniform bins (quantiles) and WOE measures were calculated. Backward logistic regression was used. In the final scorecard, six inputs were selected with an AUROC of 0.745.

Use was made of SMOTE to create 100 percent synthetic cases. Similarly the same inputs were discretised (five uniform bins), WOE measures calculated and a backward logistic regression scorecard fitted on the SMOTEd dataset. Again, six significant inputs were selected with an AUROC of 0.819.

The SMOTEd dataset includes both the original and synthetic observations. It is therefore necessary to use the scorecard constructed on the SMOTEd dataset to score the original unSMOTEd dataset to assess the performance of SMOTE. Using the scorecard from the SMOTEd dataset, an AUROC of 0.753 was obtained on the original unSMOTEd dataset, which shows a marginal improvement in predictive performance using the SMOTE technique. The results of the preliminary investigation are summarised in Table 4.1.

Scorecard	Number of records	Inputs in scorecard	AUROC
UnSMOTEd original dataset (WOE)	1875	6	0.745
SMOTE dataset (WOE)	2250	6	0.819
UnSMOTEd original dataset with SMOTE (WOE) scorecard applied	1875	6	0.753

Table 4.1: Summary of results of SMOTE evaluation

The preliminary investigation shows that the discriminatory power of the fraud risk scorecard improved when using SMOTE. The SMOTE evaluation shows promising results. Although the investigation used only a few inputs (seven original and six significant inputs), using SMOTE improved the performance of the scorecard.

In an environment where scorecards are developed for prediction, it is recommended that SMOTE be considered to improve the predictive performance of the scorecards.

Note that SMOTE was not applied in the empirical case studies as the WOE measures were used to explore the original distribution of the inputs (applying SMOTE add synthetic observations to the data).

In a production environment, the use of SMOTE may improve the discriminatory power of the fraud risk scorecards.

4.3 CASE STUDY ONE: SHORT-TERM INSURANCE DATA

The first case study uses a dataset derived from short-term insurance data in the South Africa insurance industry. The dataset is used to illustrate how a SAAS is utilised for fraud detection. The empirical investigation evaluates the hypothesis that data residing in large databases of an insurance organisation can be utilised to detect fraud. The data contain information pertaining to the policyholder, agent and the claims history. It includes a textual data field, namely the claim narrative, which contains additional textual information related to the claim. The approach of putting a SAAS in place, as described in Chapter 2, was followed as far as possible for a sub-section of policies, namely the motor sector, keeping in mind the constraint of data quality.

4.3.1 Organisational objectives

The most costly and urgent types of crime were identified (motor insurance).

In addition, the following organisational objectives were identified:

- The cost and occurrences of insurance fraud reached unacceptable levels and needed to be lowered.
- Once suspicious activity were identified, investigations took too long due to a lack of a centralised data platform, while a large proportion of investigated cases were unfounded, due to ineffective fraud detection rules. The investigation period needed to be shortened and fraud detection rules improved.

4.3.2 Collect the data

Data from the data warehouse and other internal sources (claims, policies, customers and agents) were combined with data about investigated fraud cases (see Figure 4.1).

Based on the available data, fraud risk scorecards were developed for two types of entities: policyholders and agents. Steps were followed to address data quality issues, like the exclusion of variables with too many missing values (more than 50%) and inconsistent values. The datasets from the data warehouse and other sources were combined to form analytical base tables. A binary fraud indicator was added to identify investigated cases where fraudulent activity was found. Sets of univariate inputs (identified from diagnostic fraud indicators and the literature) were calculated and added to the dataset. Depending on data availability, the inputs were grouped into the following sub-categories for the two entities:

- Demographical data (characteristics of the entities, e.g. age, occupation etc.)
- Static data (inputs available at the time of claim)
- Dynamic data (behavioural characteristics of the entity derived from historic data)
- Text data (natural language free-form text fields containing non-numeric data)
- Network data (data about the network an entity forms part of)

The datasets and sub-categories enabled the study to investigate the incremental value of including vectors from the input dimension into the scorecard during the scorecard development process for the two types of entities.

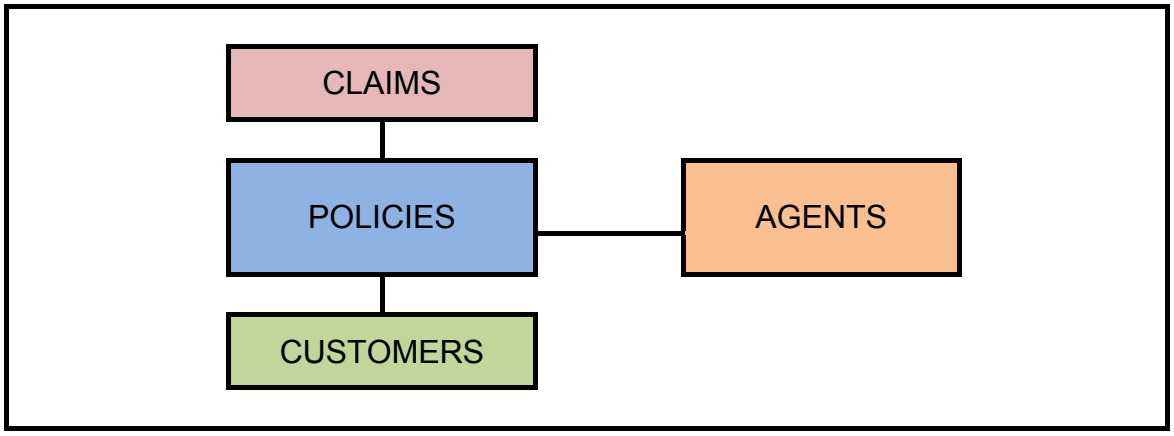


Figure 4.1: Combined datasets of an insurance organisation

4.3.2.1 Policyholder dataset

The dataset of known fraudulent claims at policyholder level was unbalanced (2% of claims received were found to be fraudulent). Random under sampling of the majority class was followed to create an analytical base table with 80 percent non-events (non-fraudulent) and 20 percent events (fraudulent). The random under sampling increased the proportion of the more important minority class for scorecard development and contained a more realistic proportion of the minority class, corresponding to the realistic fraud rates as mentioned in the literature. A wide set of anomaly indicators and other risk factors were added to the analytical base tables.

4.3.2.2 Agent dataset

The dataset of known fraudulent agents was also unbalanced (less than 5% of known agents were found to be involved in suspicious activity). Following the same rationale and random under sampling of the minority class, the final analytical base table for scorecard development contained 10 percent of fraudulent agents, using all of the suspicious observations from the all-fraudulent minority agent data.

4.3.3 Segmentation

As mentioned, the case study focused on motor claims. In other words, fraud risk scorecards were developed for the motor segment only. The following inputs were identified as candidates for further segmentation for the policyholder scorecard:

- Market Segment
- Product Type
- Premium Type
- Loss Class

4.3.4 Identification and evaluation of the risk factors

4.3.4.1 Policyholder risk factors

A wide set of risk factors were made available for analysis. WOE measures and IVs were used to evaluate the relevancy of the risk factors. Several risk factors showed a significant relationship with the outcome variable: fraudulent activity, based on the interpretation of the IV (*cf.* Section 2.3.1).

For illustration purposes, examples of several of the significant inputs are provided. The significant inputs are grouped by input category.

Demographical inputs that show a significant relationship with fraudulent activity are as follows (and summarised in Table 4.2):

- The region of home address of main policyholder (medium IV of 0.288)
- The occupation of main policyholder (medium IV of 0.101)
- The gender of main policyholder (weak IV of 0.072)
- The age of the main policyholder (weak IV of 0.02).

Variable	Information Value (IV)
Region	0.288
Occupation of main policyholder	0.101
Gender of main policyholder	0.072
Age of main policyholder	0.020

Table 4.2: Evaluation of the demographical risk factors and its corresponding Information Values (IVs)

The following inputs are examples of static risk factors (as summarised in Table 4.3):

- The claim amount larger than the 95 percentile by loss class (medium IV of 0.241)

- The time since inception to claim (medium IV of 0.112)
- The product type (medium IV of 0.101)
- The claim reason (weak IV of 0.044)
- The vacation indicator (unpredictive IV of 0.003).

Variable	Information Value (IV)
Claim amount outlier	0.241
Time since inception	0.112
Product type	0.101
Claim reason	0.044
Vacation indicator	0.003

Table 4.3: Evaluation of the static data risk factors and its corresponding Information Values (IVs)

The following dynamic inputs (based on historic data) displayed a statistically significant relationship with the fraudulent outcome (as summarised in Table 4.4):

- The total amount claimed on the policy (strong IV of 0.73)
- The number of claims on the policy (medium IV of 0.17)
- Whether or not the policy had previous claims within a short period after inception (weak IV of 0.09)
- Network density (weak IV of 0.063)
- The time period between the current claim and previous claim (weak IV of 0.06)

Variable	Information Value (IV)
Total claim amount by policy	0.73
Total claim frequency by policy	0.17
Previous claims shortly after inception	0.09
Network density	0.063
Time since last claim	0.06

Table 4.4: Evaluation of the dynamic data risk factors and its corresponding Information Values (IVs)

Using text mining algorithms and synonym lists, the following concepts were extracted from the claim narrative as roll-up terms and used in subsequent analysis:

- Hi-jack indicator
- Fire indicator
- Write-off indicator
- Lightning indicator
- Registration number of the vehicle

Using the registration number of the vehicle, extracted from the claim narrative, further summary information was created including the number of previous claims using the same registration details.

By using text mining algorithms, in particular synonym lists (*cf.* Table 3.2) and roll-up terms; concepts were extracted from the textual data field, namely the claim narrative. The following roll up terms showed a statistically significant relationship with the outcome variable: fraudulent activity, based on significant IVs (as summarised in Table 4.5):

- The roll up term depicting fire (medium IV of 0.23)
- The roll up term depicting hi-jack (medium IV of 0.19)
- The sum of claims using similar registration details (unpredictive IV of 0.008).

Variable	Information Value (IV)
Roll up 1 term: fire	0.23
Roll up 2: hi-jack	0.19
Sum of same registration details	0.008

Table 4.5: Evaluation of the text data risk factors and its corresponding Information Values (IVs)

Pearson correlation tests and the VARCLUS procedure in SAS were used as input redundancy evaluation methods (SAS Institute, 2010). The VARCLUS procedure is based on principal component analysis of the numeric inputs. Inputs are grouped together based on a threshold set for the second eigenvalue. Once the variables are grouped into clusters, variables with a low $1\text{-}R^2$ ratio will be good representatives of the cluster. An extract of the variable clustering results are provided in Table 4.6. The analysis identified eight clusters. The representative of each cluster can be decided on by using the low $1\text{-}R^2$ ratio or business considerations.

Cluster	Variable	OwnCluster	NextClosest	RSquareRatio
Cluster 1	Time to expiry of policy	0.571352796	0.350501307	0.65996623
	Young policy claims indicator	0.92335047	0.199824652	0.095790917
	Claim frequency on young policy	0.89464774	0.213611315	0.133969704
Cluster 2	Total loss amount to date	0.947569108	0.533957214	0.112502315
	Broker claim frequency	0.928865184	0.201199954	0.089052092
	Broker total loss amount	0.92923777	0.205810304	0.08909991
	Claim amount	0.882126334	0.599595899	0.29438676
Cluster 3	Time since inception	0.891649576	0.335466659	0.163047386
	Claim frequency	0.791738201	0.320323684	0.306413207
	Duration of broker relationship	0.54125524	0.177180889	0.557528082
Cluster 5	Days since incident to reported	0.48674009	0.067515658	0.550422014
	Day of week of incident	0.933063455	0.192610627	0.082904912
	Day of month of incident	0.933063455	0.192610627	0.082904912
Cluster 6	Previous suspicious claim indicator	0.421993363	0.299563886	0.825209645
	Close to expiry indicator	0.671772176	0.22341868	0.422657378
	Days since previous claim	0.384678844	0.101849957	0.685098398
	Suspicious broker indicator	0.407465786	0.182628454	0.724926402
Cluster 7	Age of main policy holder	0.830058526	0.207554813	0.214452024
	Customer or company indicator	0.737672529	0.169876452	0.316010155
Cluster 8	Loss amount outlier trigger	0.865988512	0.355995995	0.208091079
	Average claim amount	0.807755813	0.382345019	0.311248501
	Broker guarantee indicator	0.496153152	0.295165944	0.714844642

Table 4.6: Assignment of numeric variables to clusters based on variable clustering

4.3.4.2 Agent risk factors

As a second entity level scorecard, agent level analysis was performed. The agents fulfil an important role as intermediary between the insurance organisation and its policyholders. The interaction provides agents with an opportunity to learn the strengths and weaknesses of the insurance systems with which they interact on a daily basis and an opportunity to commit fraud. Based on the investigated fraud cases in the case study, several agents were either solely responsible for fraud or worked together with other entities to defraud the insurance organisation. These agents were flagged with an agent-fraud indicator. Diagnostic fraud risk factors were added to the analytical base table, together with summary information on the behavioural characteristics.

Table 4.7 shows some of the significant risk factors at agent level with their corresponding IVs.

Some of the statistically significant risk factors at agent level are

- Number of claims (strong IV of 0.63)
- Type of agent (strong IV of 0.62)
- Length of relationship with the organisation (strong IV of 0.55)
- Total amount paid to agent in the last month (strong IV of 0.39)
- Geographic region (medium IV of 0.25)

Variable	Information Value (IV)
Number of claims	0.63
Type of agent	0.62
Relationship with organisation	0.55
Total amount paid	0.39
Geographic region	0.25

Table 4.7: Risk factors of the agent-level scorecard

4.3.4.3 Examples of risk factor identification and evaluation

Examples of the effects and WOE measures of some of the statistically significant risk factors at policyholder level are provided and described in more detail.

The overall frequency distribution of each input across the bins is provided in each graph. Note that the WOE measures make use of the relative class frequencies of the bins, in other words, the overall frequency distribution does not determine the WOE measures. The overall frequency distribution gives the reader an idea of the distribution of the data across the bins, whilst the WOE measures indicate the fraud risk of each bin. High negative values of the WOE are associated with high fraud risk and high positive values of the WOE are associated with low fraud risk.

1. Risk factor of the time since inception to claim

Specialist investigators refer to the time shortly after inception of a policy as the ‘green business’ period. Claims within the ‘green business’ period have higher fraud risk according to the judgemental knowledge of the subject matter experts. The risk factor was evaluated using WOE measures and its IV. Based on the data, Figure 4.2 depicts the WOE graph for the time since inception to claim risk factor while Table 4.8 indicates the values of the WOE measures and corresponding IV of 0.112 (medium).

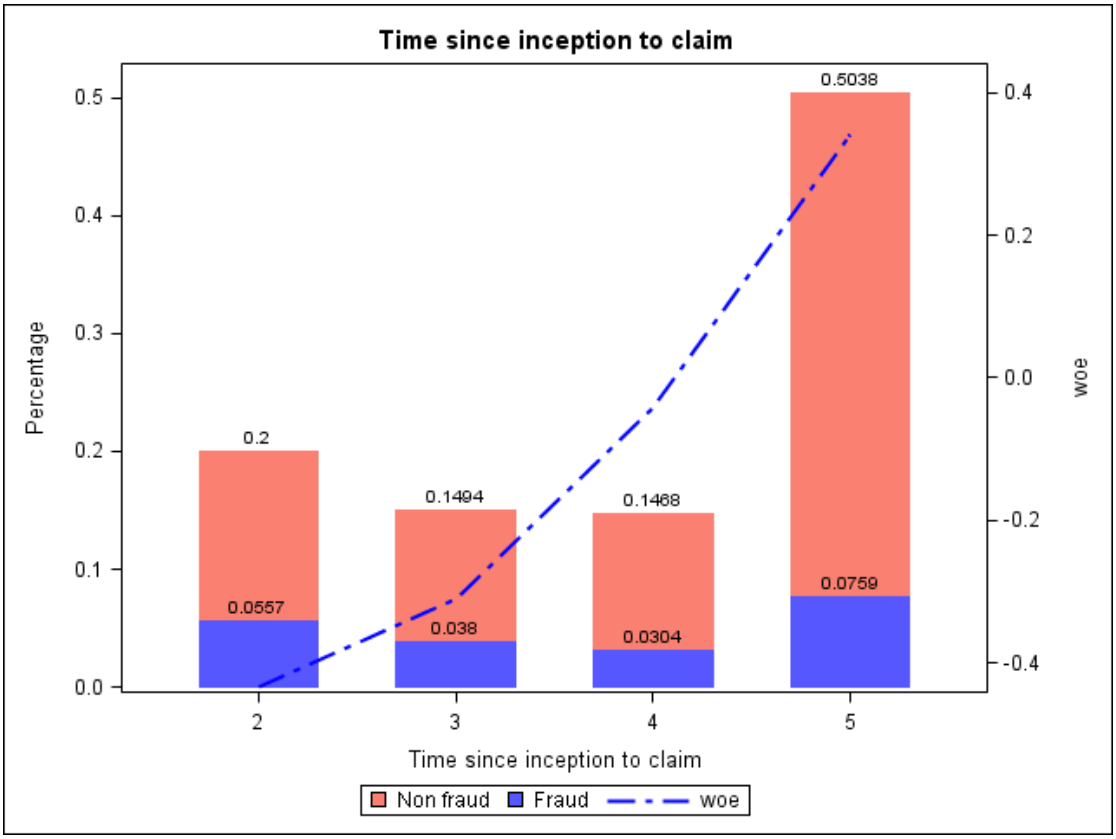


Figure 4.2: Weights-of-evidence (WOE) measures of the time since inception to claim input

Attribute	Total %	Event %	Fraud Distribution	Non-fraud Distribution	WOE
Bin 1	0.200	0.056	0.278	0.180	-0.434
Bin 2	0.149	0.038	0.190	0.139	-0.310
Bin 3	0.147	0.030	0.152	0.146	-0.043
Bin 4	0.504	0.076	0.380	0.535	0.342
Information Value (IV)					0.112

Table 4.8: Frequency counts and weights-of-evidence (WOE) values of the time to claim input

From the WOE graph in Figure 4.2, it is evident that high fraud risk exists when the claim occurred shortly after inception of policy. The short-term insurance data confirms the risk factor. The WOE graph demonstrates higher fraud risk for younger policies, as observed for the first bin, with the fraud risk decreasing gradually for claims later on in the policy life time.

2. Risk factor of the vacation indicator

Some studies found an increase in fraudulent activity during weekends and vacation periods (Phua *et al.*, 2004). Figure 4.3 depicts the WOE graph for the vacation indicator risk factor while Table 4.9 indicates the values of the WOE measures and corresponding IV of 0.003 (unpredictive).

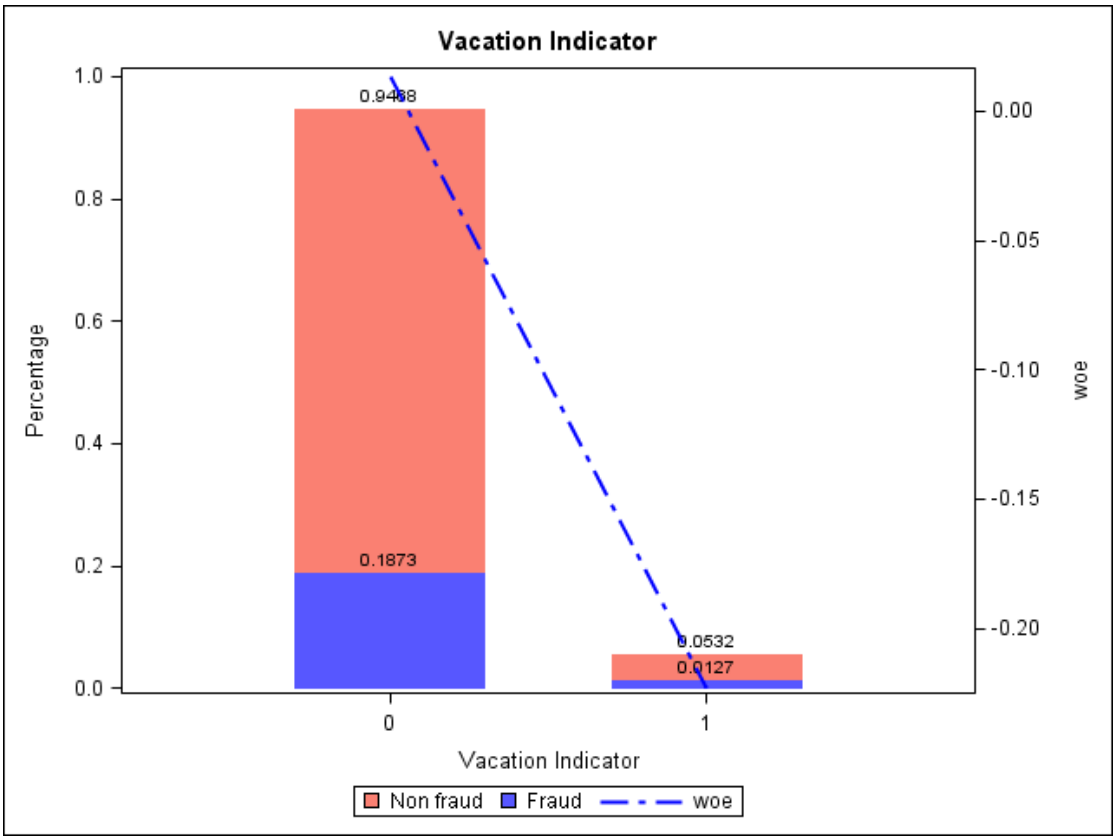


Figure 4.3: Weights-of–evidence (WOE) measures for the vacation indicator

Bin	Attribute	Total	Event	Fraud Distribution	Non-fraud Distribution	WOE
1	Vacation ind=0	0.947	0.187	0.937	0.949	0.013
2	Vacation ind=1	0.053	0.013	0.063	0.051	-0.223
Information Value (IV)						0.003

Table 4.9: Frequency counts and weights-of-evidence (WOE) values of the two bins of the vacation indicator variable

In the case study, the WOE measures of the binary vacation indicator demonstrate a slightly higher fraud risk for incidents during a vacation period (like Christmas). It should be noted that in terms of motor claims, and in general, more incidents occur during vacation periods, due to higher volume of traffic for example. However, the WOE measures demonstrate that more fraudulent events were observed during the vacation periods proportional to the frequency of claims. Based on the statistical analysis, namely small IV, the vacation indicator should not be included in the scorecard. However, due to business considerations, the input was included.

3. Risk factor of the sum of claims using similar registration numbers

The registration number of the vehicles were extracted from the textual data field: claim narrative. The sum of claims using similar registration numbers was evaluated as a risk factor. Figure 4.4 depicts the WOE graph for sum of claims using similar registration numbers, while Table 4.10 indicates the values of WOE measures and corresponding IV of 0.008 (unpredictive).

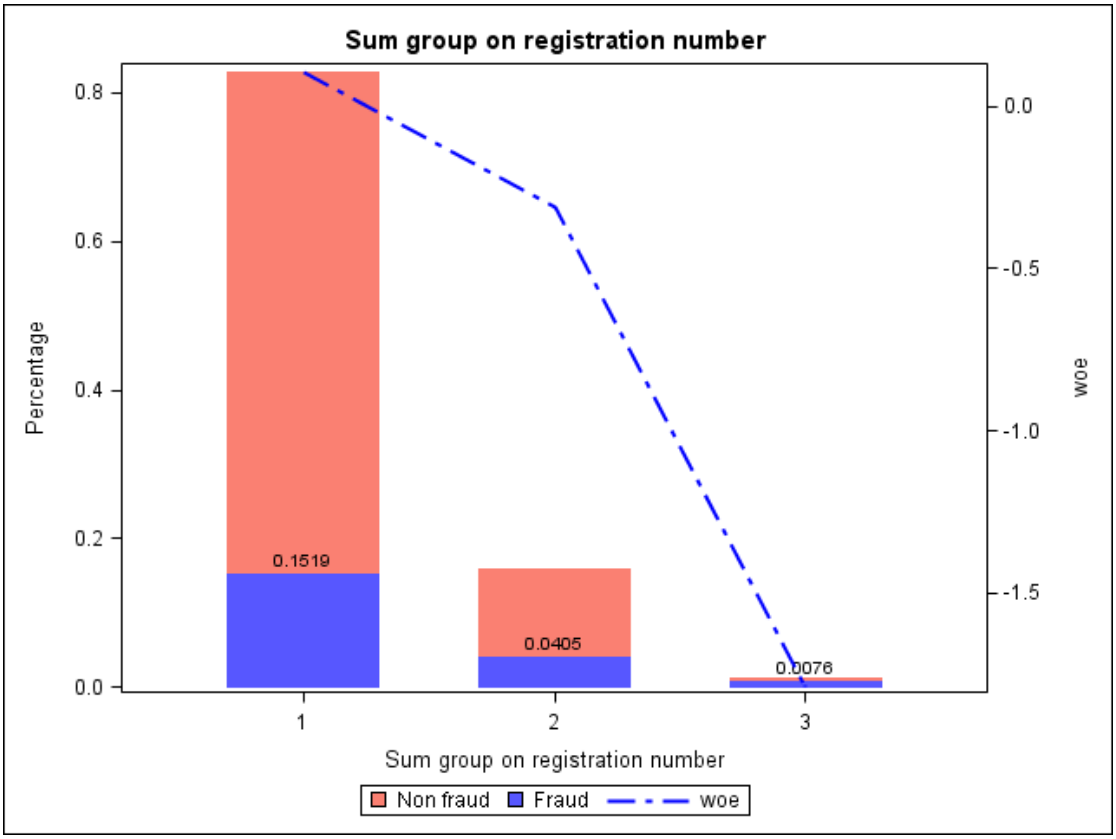


Figure 4.4: Weights-of-evidence (WOE) measures for the sum of claims using the same registration details

Bin	Attribute	Total %	Event %	Fraud Distribution	Non-fraud Distribution	WOE
1	Sum group 1	0.828	0.152	0.759	0.845	0.107
2	Sum group 2	0.159	0.040	0.203	0.149	-0.309
3	Sum group 3	0.013	0.008	0.038	0.006	-1.792
Information Value (IV)						0.008

Table 4.10: Frequency counts and weights-of-evidence (WOE) values of the three bins of the sum on registration number variable

The WOE graph indicates that slightly higher risk of fraud is observed for larger frequencies of claims using similar registration numbers. The WOE measures indicate that perpetrators make use of a loop hole in the claims information capturing process: the registration number is captured as free-

form text making it difficult to measure and monitor the frequency of claims using the same details. Based on the statistical analysis, namely a small IV, the similar registration numbers input should not be included in the scorecard. However, due to business considerations, the input was included.

4. Risk factor of the broker density

The study measured the amount of activity by brokers (submitted and cancelled claims) in their networks. Figure 4.5 depicts the WOE graph for broker density, while Table 4.11 indicates the values of WOE measures and corresponding IV of 0.063 (weak).

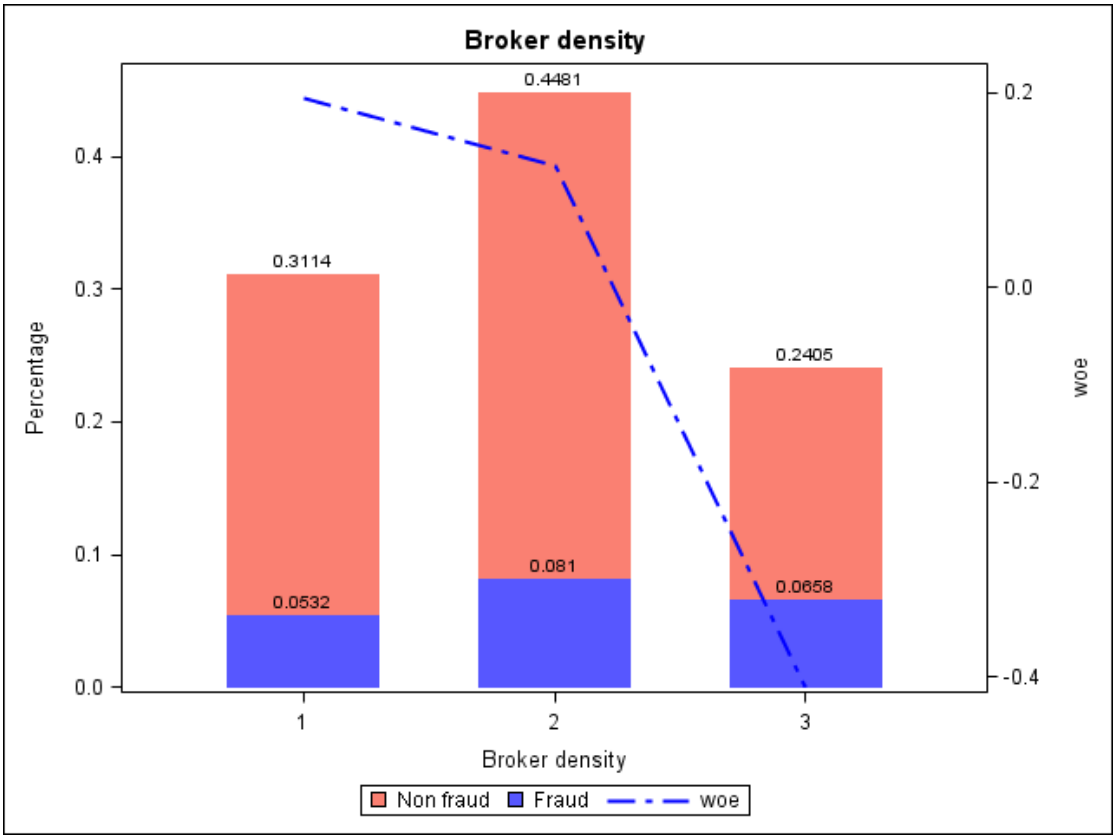


Figure 4.5: Weights-of-evidence (WOE) measures of the broker density

Bin	Attribute	Total %	Event %	Fraud Distribution	Non-fraud Distribution	WOE
1	Low activity	0.311	0.053	0.266	0.323	0.194
2	Medium activity	0.448	0.081	0.405	0.459	0.125
3	High activity	0.241	0.066	0.329	0.218	-0.410
Information Value (IV)						0.063

Table 4.11: Frequency counts and weights-of-evidence (WOE) values of the three bins of the network density variable

The WOE measures of the broker density show a statistically significant relationship between an increase in the density of a broker's network (more activity) and fraud risk. That is, more active networks contain more fraudulent activity.

5. Risk factor of the total claim amount

The total claim amount is the sum of all claims on the policy up to date. The total claim amount was evaluated as a risk factor. Figure 4.6 depicts the WOE graph for the total claim amount, while Table 4.12 indicates the values of WOE measures and corresponding IV of 0.73 (strong).

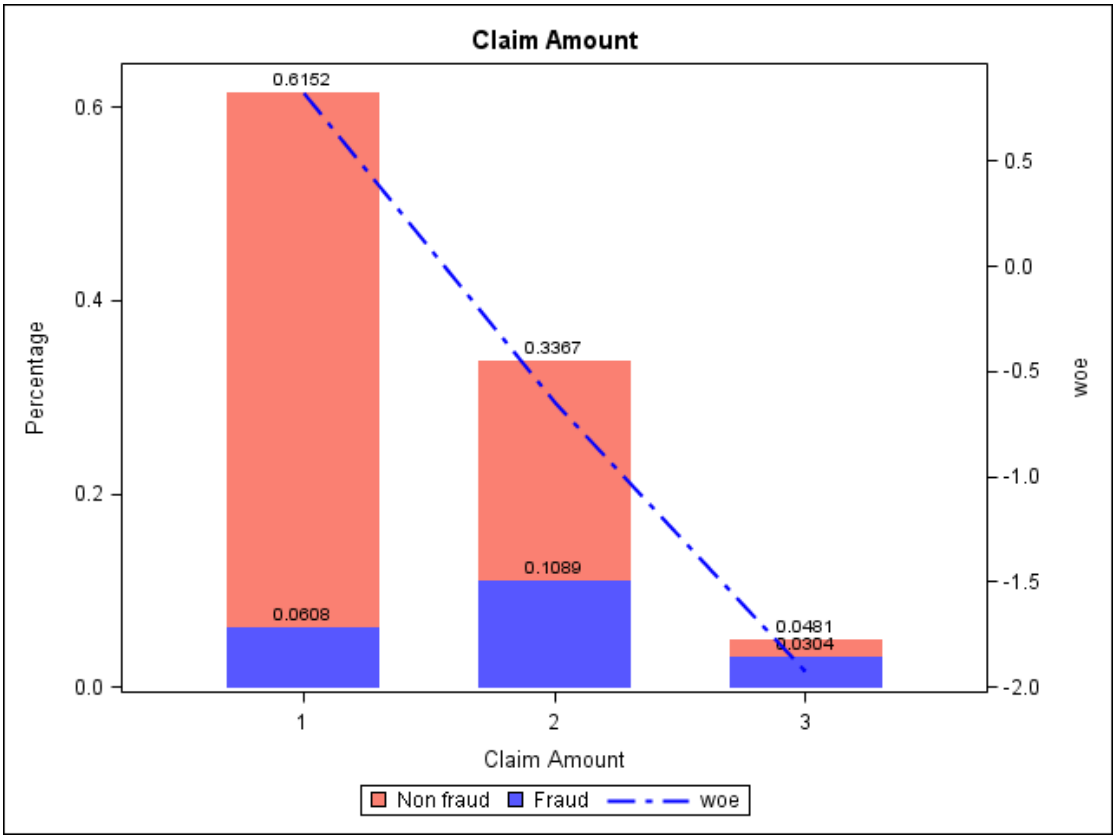


Figure 4.6: Weights-of-evidence (WOE) measures of the claim amount

Bin	Attribute	Total %	Event %	Fraud Distribution	Non-fraud Distribution	WOE
1	Small total	0.615	0.061	0.304	0.693	0.825
2	Medium total	0.337	0.109	0.544	0.285	-0.648
3	Large total	0.048	0.030	0.152	0.022	-1.925
Information Value (IV)						0.73

Table 4.12: Frequency counts and weights-of-evidence (WOE) values of the three bins of the claim amount input

The WOE measures display that higher risk of fraud is observed for larger accumulated losses.

4.3.5 Fit the fraud risk scorecards

4.3.5.1 Policyholder fraud risk scorecard

The final policyholder scorecard used backward logistic regression and the following variables are included in the scorecard (all with significant p-values, at a significance level of $p < 0.05$):

- Total claim amount on policy
- Policy premium
- Days since inception to claim (time to claim)
- Days to reported
- Roll up term: fire
- Region of home address of main policy holder

The all-negative parameter estimates of the WOE inputs are provided in Table 4.13.

Input	Parameter Estimate	p value
Total Claim Amount	-0.7645	<0.0001
Policy Premium	-0.7394	<0.0001
Time to claim	-0.7411	0.0329
Days to Reported	-0.9081	0.0395
Roll up term: fire	-0.7247	0.0073
Region	-0.7085	0.0256

Table 4.13: Analysis of the maximum likelihood estimates of the policyholder scorecard

Using the parameter estimates, WOE and scorecard scaling formulae, a scorecard may be constructed. Scorecards may ease the interpretation of the fraud detection assessment; however the output from the logistic regression is a probability score, which in itself may be used. In the case of a probability score, the threshold would be set as a maximum: entities with a probability score larger than the threshold would be flagged for investigation.

4.3.5.2 Agent fraud risk scorecard

The final agent scorecard uses backward logistic regression and the following inputs are included in the scorecard (all with significant p-values, at a significance level of $p < 0.05$):

- Geographic region of the agent
- Type of agent
- Total paid out in the previous month
- Length of relationship

The all-negative parameter estimates of the WOE inputs are provided in Table 4.14.

Parameter	Estimate	p value
Geographic region	-0.6794	0.0003
Type of agent	-1.4391	0.0129
Total amount paid last month	-1.1480	0.0013
Relationship with organisation	-0.7710	0.0182

Table 4.14: Analysis of the maximum likelihood estimates of the agent scorecard

4.3.6 Validation

4.3.6.1 Policyholder fraud risk scorecard

The policyholder scorecard was validated using the performance measures described in Chapter 2. Intuitively, better performance is expected by including more risk factors, so the improvement in performance of the scorecard across the range of input vectors is expected. As can be seen by evaluating the ROC curve depicted in Figure 4.7 (with corresponding AUROC and AR measures in Table 4.15), using only demographical information provides some lift (AUROC of 0.387). However, by including more input variables about the claims and policy history, much better lift is observed in terms of the ROC curve and its corresponding AUROC of 0.85. The addition

of text data lead to a further improvement in discriminatory power (AUROC to 0.87).

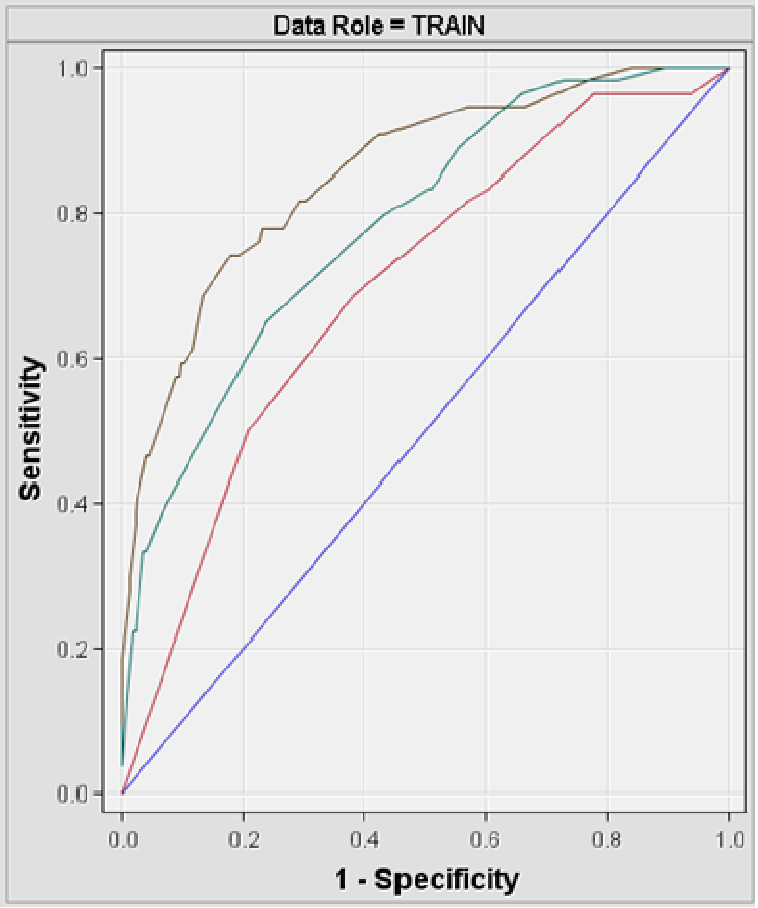


Figure 4.7: Receiver Operating Curve (ROC) of the policyholder scorecard (demographical, static and dynamic data)

Scorecard	AUROC	AR
Demographical data	0.69	0.387
Adding static data (claim and policy information)	0.79	0.584
Adding dynamic data (claims and policy history)	0.85	0.704
Adding text data	0.87	0.71

Table 4.15: Performance measures of the policyholder scorecard

	Predicted Positives	Predicted Negatives
Actual Positives	42	37
Actual Negatives	15	301

Table 4.16: Confusion matrix for the policyholder scorecard

As an illustration, for a subset of claims used for scorecard testing, the policyholder fraud risk scorecard identified 42 out of 79 fraudulent policyholders (53%) accurately, whilst 42 out of 57 (74%) of the number of policyholders flagged for fraud where fraudulent (confusion matrix in Table 4.16).

4.3.6.2 Agent fraud risk scorecard

For the agent fraud-risk scorecard, the AUROC on the validation dataset is 0.86, whilst the AR is 0.72 (see Table 4.17). Table 4.18 shows the confusion matrix on a subset of claims used for scorecard testing. It indicates that 18 out of 32 (56%) fraudulent agents were correctly classified by the risk scorecard. Of the 20 agents flagged for investigation by the risk scorecard, 18 were found to be fraudulent (90%).

Scorecard	AUROC	AR
Training Dataset	0.87	0.74
Validation Dataset	0.86	0.72

Table 4.17: Perfomance results for the agent scorecard

	Predicted Positives	Predicted Negatives
Actual Positives	18	14
Actual Negatives	2	242

Table 4.18: Confusion matrix for agent fraud scorecard

4.3.7 Summary of the results

The case study illustrates how a SAAS is applied to South African short-term insurance data.

The organisational objectives were identified: to reduce payments to fraudulent claims and to identify risk factors associated with fraudulent activity.

A sample of claims data was made available for a sub-segment of the business, namely motor claims.

The risk factors pertaining to two types of entities were identified from the data, namely policyholders and agents. Techniques like anomaly detection, text mining and relational data, and expert knowledge were used to identify risk factors. An example of an anomaly is the large claim amount indicator by loss class. An example of the use of relational data is the use of the broker density. An example of the use of expert knowledge is the time since inception to claim input. Specialist investigators refer to the period shortly after inception as the 'green business period'. Based on the data availability and quality, a wide set of risk factors were made available for analysis for the two types of entities.

The WOE measures and corresponding IVs were used to identify relevant risk factors (univariate analysis). The variable clustering procedure was used to evaluate the redundancy of the inputs. To determine flexible split points for the WOE measures, a combination of statistical techniques and business considerations were used.

Fraud risk scorecards were constructed using WOE measures and backward logistic regression (multivariate analysis). The logistic regression scorecards explain the combined effect of the statistically significant risk factors to estimate a fraud risk probability using demographic, static, dynamic, textual and relational risk factors.

The scorecard at policyholder level using all inputs obtained an AUROC of 0.87. The AUROC value is reasonable.

The scorecard at agent level using all inputs obtained an AUROC of 0.86 on the validation dataset. The AUROC value is reasonable.

Using business input, the cut-off across the posterior probability range may be determined based on the maximum profit or minimum cost. The cut-off may also be determined based on business input like the capacity of the investigations unit, for example.

The empirical case study on short-term insurance data demonstrates how the WOE measures (and corresponding IVs) are used to identify risk factors for fraud detection. The logistic regression-based scorecards explain the combined effect of the statistically significant risk factors to estimate a fraud risk probability.

4.4 CASE STUDY TWO: LIFE INSURANCE DATA

According to Stokes (2010), death and funeral policies remain the largest contributor of fraudulent insurance claims in South Africa.

The second case study uses a sample of claims from a life insurance organisation in South Africa. For confidentiality purposes, personal information was removed from the dataset. The empirical investigation evaluates the hypothesis that a SAAS as outlined in Chapter 2 can identify fraudulent activity successfully. The performance criteria of what constitutes successful fraud detection include a significant AUROC, visually illustrated using the ROC curve.

4.4.1 Organisational Objectives

The most costly and urgent types of crime were identified (fraudulent claims on funeral products).

In addition, the following organisational objectives were identified:

- The cost and occurrences of insurance fraud reached unacceptable levels and needed to be lowered.
- The risk factors associated with fraudulent activity needed to be identified.

4.4.2 Collect the data

The data contains information pertaining to the claimant, agent, and the claims history. Unfortunately, no textual data fields were available for analysis and limited relational data was available. The approach of putting a SAAS in

place, as described in Chapter 2, was followed as far as possible on a sample of claims on funeral products, keeping in mind the constraint of data quality.

Policyholder data was combined with data about claims. A binary fraud indicator was added to identify investigated claims where fraudulent activity was found. Sets of univariate inputs (identified from diagnostic fraud indicators and the literature) were calculated and added to the dataset. The available inputs were grouped according to the following sub-categories:

- Demographical data (characteristics of the entity, e.g. age, occupation etc.)
- Static data (inputs available at the time of claim)
- Dynamic data (behavioural characteristics of the entity derived from historic data)

The dataset and sub-categories enabled the study to investigate the incremental value of including vectors from the input dimension into the scorecard during the scorecard development process. Several data quality and data improvement steps were taken to improve the usefulness of the data. For example, postcodes were merged with a data table from Geonames (2011) to identify the geographic region.

The dataset of known fraudulent claims at claimant level was unbalanced (5.6% of all claims received were found to be fraudulent).

4.4.3 Segmentation

The case study focused on funeral products. In other words, fraud risk scorecards were developed for the funeral product segment only. The following variables were identified as candidates for further segmentation:

- Product Type
- Group scheme / Individual policies
- Market Segment
- Premium Type

4.4.4 Identification and evaluation of the risk factors

For the fitting of the scorecard at claimant level, the inputs were split into groups of input categories, namely

- Demographical inputs
- Static claim inputs
- Dynamic inputs

The important inputs from each group of variables are given in Table 4.19. The inputs were evaluated based on its corresponding IV and WOE distribution. The numeric and categorical inputs were collapsed based on the approaches set out in Chapter 2. The bins were evaluated and adjusted based on business considerations. The WOE measures were calculated for the new levels of the inputs and used in subsequent analysis.

Table 4.19 lists a subset of the most significant inputs at the claimant level of the life insurance dataset.

Category	Variable	Information Value (IV)
Demographic	Home address region of main insured	2.7
Demographic	Occupation of main insured	0.119
Static	Time since inception to claim	1.114
Static	Claim cover ratio	0.684
Static	Cover amount	0.521
Static	Claim reason	0.264
Dynamic	Suspicious agent indicator	0.64

Table 4.19: Examples of risk factors on life insurance data

As examples, the following demographic inputs demonstrated a statistically significant relationship with the outcome variable:

- The region of the primary home address of the main insured (strong IV of 2.7)
- The occupation of the main insured (medium IV of 0.119)

The static inputs with a statistically significant relation with the outcome: fraudulent activity, were

- The time period from inception of the policy to current claim (strong IV of 1.114)
- The claim cover ratio (strong IV of 0.684)
- The cover amount (strong IV of 0.521)
- The claim reason or cause of claim (medium IV of 0.264).

The statistically significant dynamic inputs include:

- The indicator whether or not the agent was involved in suspicious activity (strong IV of 0.364).

4.4.4.1 Examples of risk factor idendification and evaluation

A few examples of the use of WOE measures and IVs for univariate data analysis on life insurance data are provided below:

1. Risk factor of the time since inception to claim

Specialist investigators refer to the time shortly after inception of a policy as the ‘green business’ period. Claims within the ‘green business’ period have higher fraud risk according to the judgemental knowledge of the subject matter experts. Based on the data, Figure 4.8 depicts the WOE graph for the time since inception to claim risk factor while Table 4.20 indicates the values of the WOE measures and corresponding IV of 1.114 (strong).

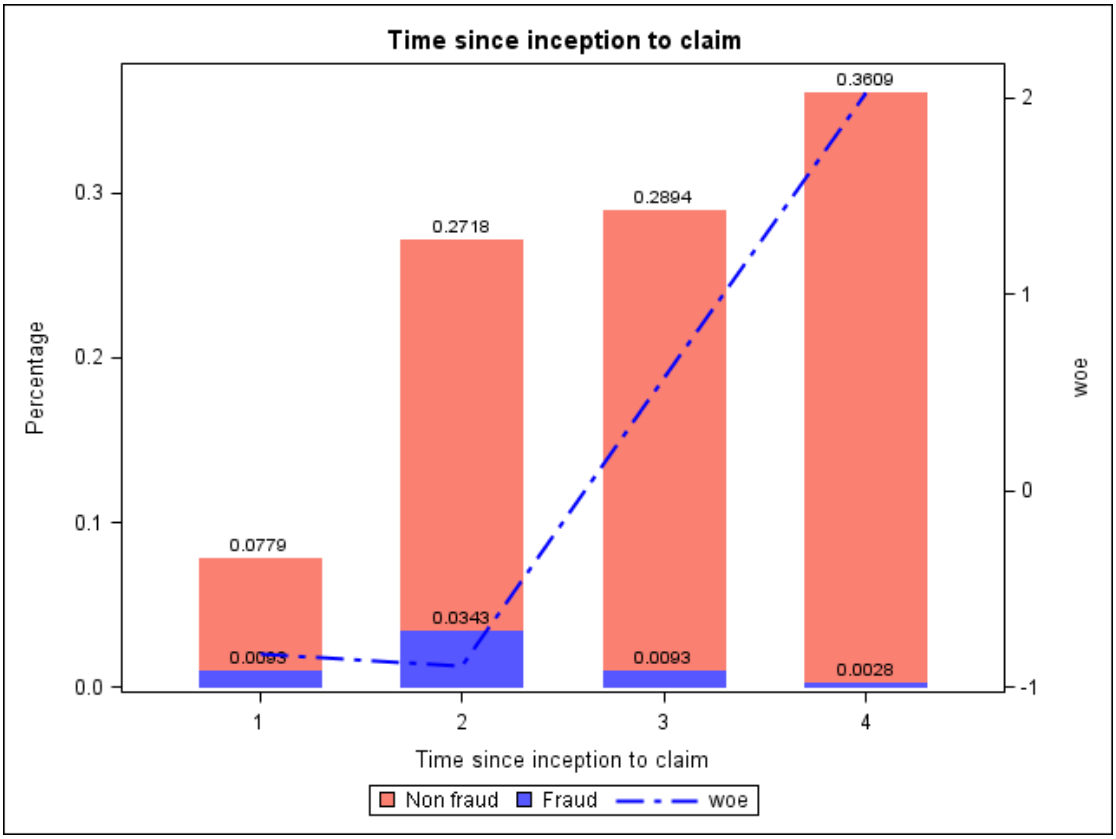


Figure 4.8: Weights-of-evidence (WOE) measures of the time since inception to claim

Bin	Attribute	Total %	Event %	Fraud Distribution	Non-fraud Distribution	WOE
1	Group 1	0.091	0.014	0.250	0.082	-1.120
2	Group 2	0.206	0.026	0.467	0.191	-0.896
3	Group 3	0.201	0.010	0.183	0.202	0.099
4	Group 4	0.502	0.006	0.100	0.525	1.659
Information Value (IV)						1.114

Table 4.20: Frequency counts and weights-of-evidence (WOE) values of the four bins of the time since inception to claim input, and corresponding Information Value (IV)

The WOE graph demonstrates higher fraud risk for younger policies, with the fraud risk decreasing gradually for claims later on in the policy lifetime.

2. Risk factor of the occupation of the policyholder

The occupation of the policyholder was evaluated as a risk factor. Based on the data, Figure 4.9 depicts the WOE graph for the occupation group risk factor while Table 4.21 indicates the values of the WOE measures and corresponding IV of 0.119 (medium).

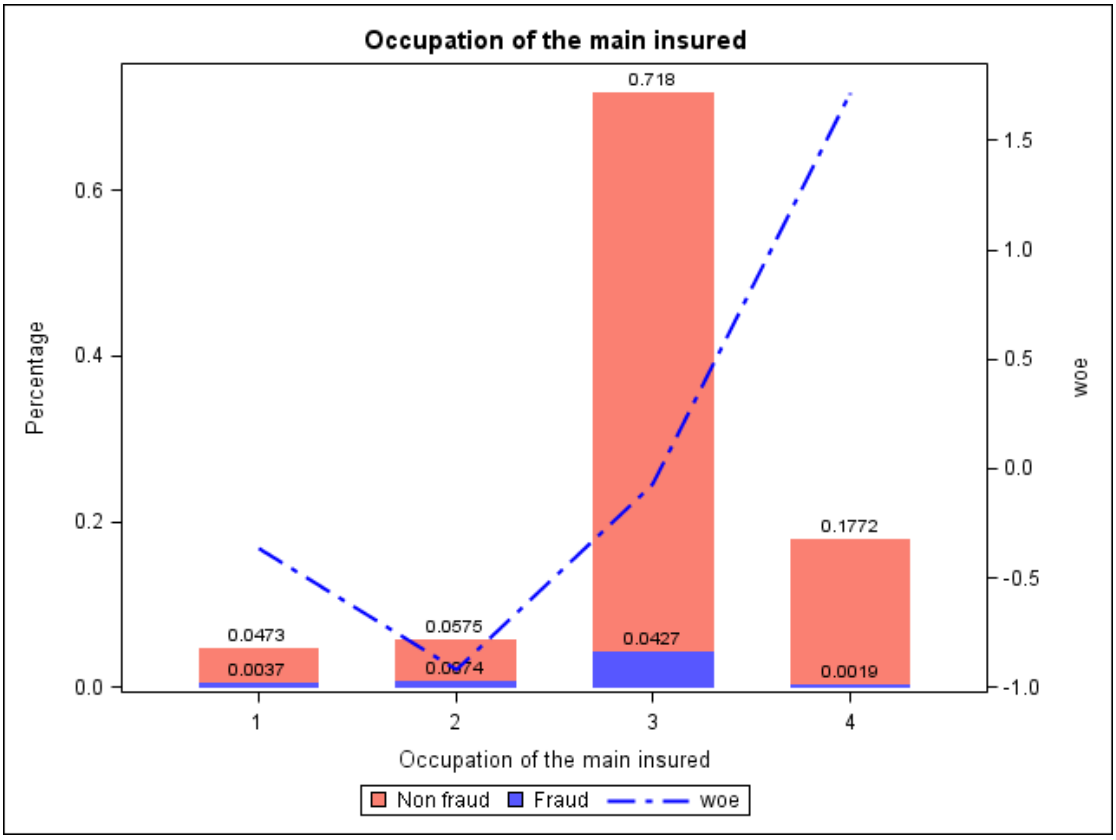


Figure 4.9: Weights-of-evidence (WOE) measures of the occupation of the policyholder

Bin	Attribute	Total %	Event %	Fraud Distribution	Non-fraud Distribution	WOE
1	Occupation group 1	0.047	0.003	0.067	0.046	-0.367
2	Occupation group 2	0.058	0.007	0.133	0.053	-0.922
3	Occupation group 3	0.718	0.043	0.767	0.715	-0.070
4	Occupation group 4	0.177	0.003	0.033	0.186	1.717
Information Value (IV)						0.119

Table 4.21: Frequency counts and weights-of-evidence (WOE) values of the four bins of the occupation variable

The WOE measures display that the fraud risk is heterogeneous across the types of occupation of the policyholders. Confidentiality prevents the publishing of the types of occupation. Occupation group 1 demonstrates the highest fraud risk, occupation groups 2 and 3 medium fraud risk, whilst occupation group 4 demonstrates low fraud risk.

3. Risk factor of the claim cover ratio

This is the ratio of the claim amount divided by the cover amount. The claim cover ratio was evaluated as a risk factor. Based on the data, Figure 4.10 depicts the WOE graph for the claim cover group risk factor while Table 4.22 indicates the values of the WOE measures and corresponding IV of 0.684 (strong).

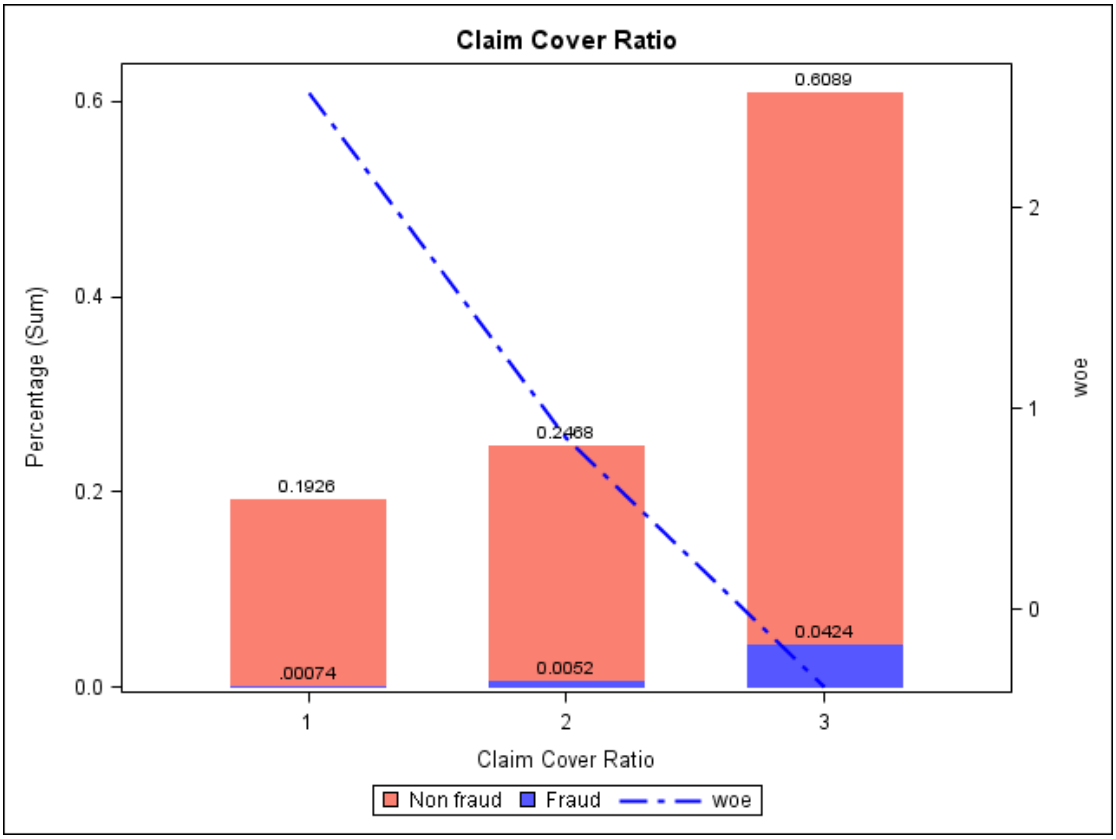


Figure 4.10: Weights-of-evidence (WOE) of the claim cover ratio groups

Bin	Attribute	Total %	Event %	Fraud Distribution	Non-fraud Distribution	WOE
1	Claim cover ratio group 1	0.183	0.002	0.015	0.193	2.573
2	Claim cover ratio group 2	0.227	0.007	0.108	0.234	0.858
3	Claim cover ratio group 3	0.59	0.047	0.877	0.573	-0.387
Information Value						0.684

Table 4.22: Frequency counts and weights-of-evidence (WOE) values of the three bins of the claim cover ratio groups

The WOE measures demonstrate that the fraud risk increases for larger values of the claim cover ratio. Claim cover ratio group 1 demonstrates low

fraud risk, claim cover ratio group 2, medium fraud risk, whilst claim cover ratio group 3 demonstrates high fraud risk.

4. Risk factor of the claim reason

The claim reason was evaluated as a risk factor. Based on the data, Figure 4.11 depicts the WOE graph for the claim reason group risk factor while Table 4.23 indicates the values of the WOE measures and corresponding IV of 0.264 (medium).

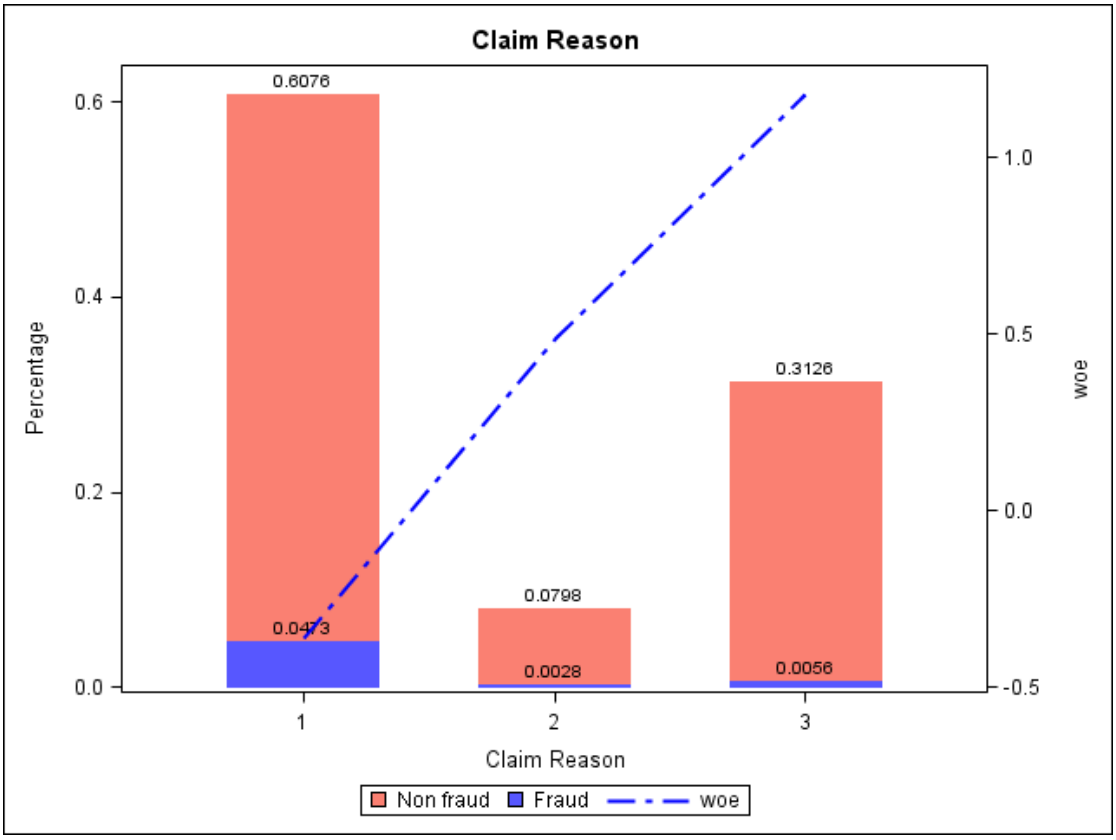


Figure 4.11: Weights-of-evidence (WOE) measures of the claim reason input

Bin	Attribute	Total %	Event %	Fraud Distribution	Non-fraud Distribution	WOE
1	Reason group 1	0.607	0.047	0.850	0.593	-0.797
2	Reason group 2	0.080	0.003	0.050	0.082	-0.206
3	Reason group 3	0.313	0.006	0.100	0.325	0.292
Information Value						0.264

Table 4.23: Frequency counts, weights-of-evidence (WOE) values of the three bins of the claim reason input, and corresponding Information Value (IV)

The WOE measures indicate that a statistically significant relationship exist between the reason of the claim and the binary outcome: fraudulent activity. The WOE measures display that specific claim reasons are more 'popular' amongst fraudsters and correspond to higher fraud risk. The claim reason group 1 displays particular high fraud risk, according to the WOE measures, whilst claim reason groups 2 and 3 displays low fraud risk.

5. Risk factor of the demographic region

The demographic region of the home address of the main policyholder was evaluated as a risk factor. Based on the data, Figure 4.12 depicts the WOE graph for the region group risk factor while Table 4.24 indicates the values of the WOE measures and corresponding IV of 2.7 (strong).

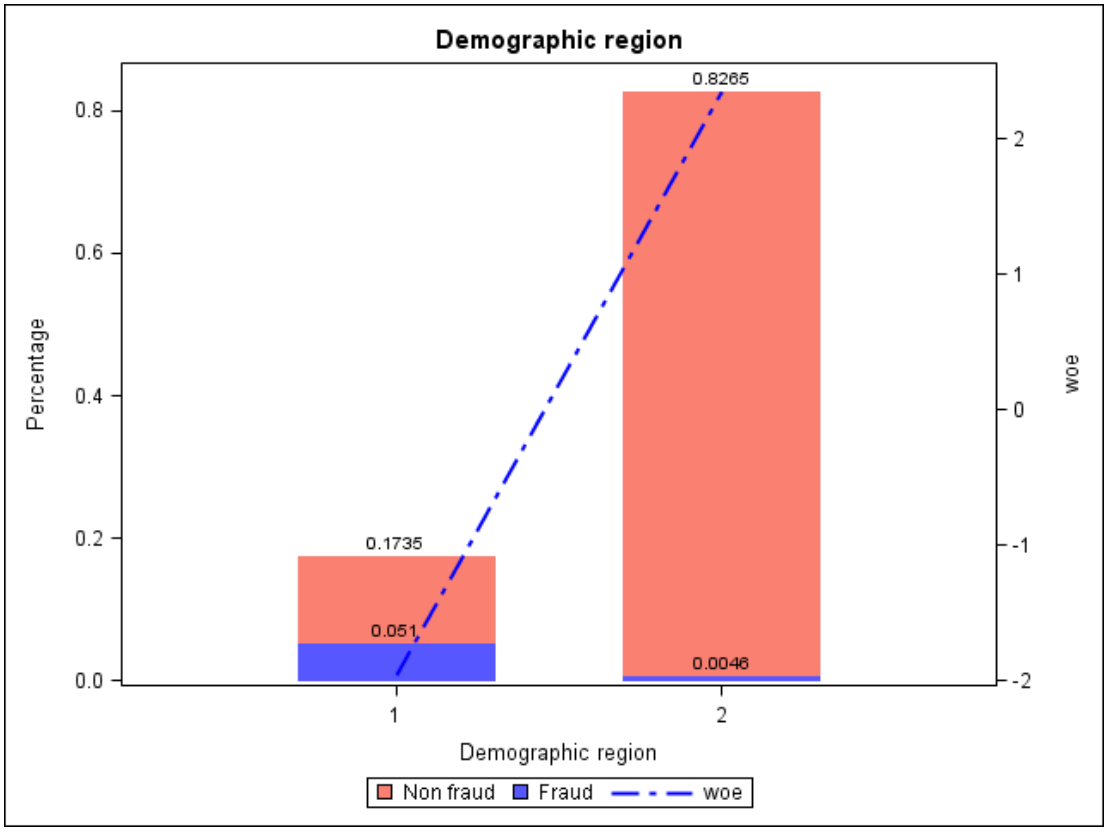


Figure 4.12: Weights-of-evidence (WOE) measures of the demographic region

Bin	Attribute	Total %	Event %	Fraud Distribution	Non-fraud Distribution	WOE
1	Region 1	0.173	0.051	0.917	0.126	-0.956
2	Region 2	0.827	0.005	0.083	0.874	2.346
Information Value (IV)						2.7

Table 4.24: Frequency counts and weights-of-evidence (WOE) values of the six bins of the demographic region variable

The IV of the region indicator is 2.7 (strong), which indicates a very high association between the region and fraudulent activity. The input should be carefully considered as its inclusion in the scorecard may overshadow the other inputs. The WOE measures indicate that a statistically significant relationship exist between the demographic region and fraudulent activity.

The WOE measures inform that the occurrence of detected fraudulent activity is region specific.

4.4.5 Fit the fraud risk scorecard

Once it was verified that no strong correlations exist between the significant inputs, backward logistic regression scorecards were fitted on the two partitions of the input data, namely the training and validation datasets, using subsets of the inputs, related to demographical inputs, static inputs and dynamic inputs, and also a final scorecard, which include all available inputs.

4.4.6 Validation

The use of demographic inputs in the risk assessment scorecard shows an AUROC of 0.576 on the validation data. The static and dynamic scorecards perform better than the demographic scorecard (on validation data: 0.756 and 0.673 respectively). A comparison of the AUROC values of the four models is provided in Table 4.25.

Category	Number of variables	AUROC (Train)	AUROC (Valid)
Demographic	2	0.609	0.576
Static	4	0.874	0.756
Dynamic	2	0.788	0.673
All	5	0.879	0.760

Table 4.25: Evaluation of the Accuracy Ratio (AR) of final scorecard

The ROC curve plots the Sensitivity and 1-Specificity values for different cut-offs or threshold values of the posterior probability. The ROC curve using validation dataset is provided for comparison in Figure 4.13. It is evident that a combination of demographic, static and dynamic inputs provides the best lift. The static scorecard performs marginally weaker than the scorecard using all inputs.

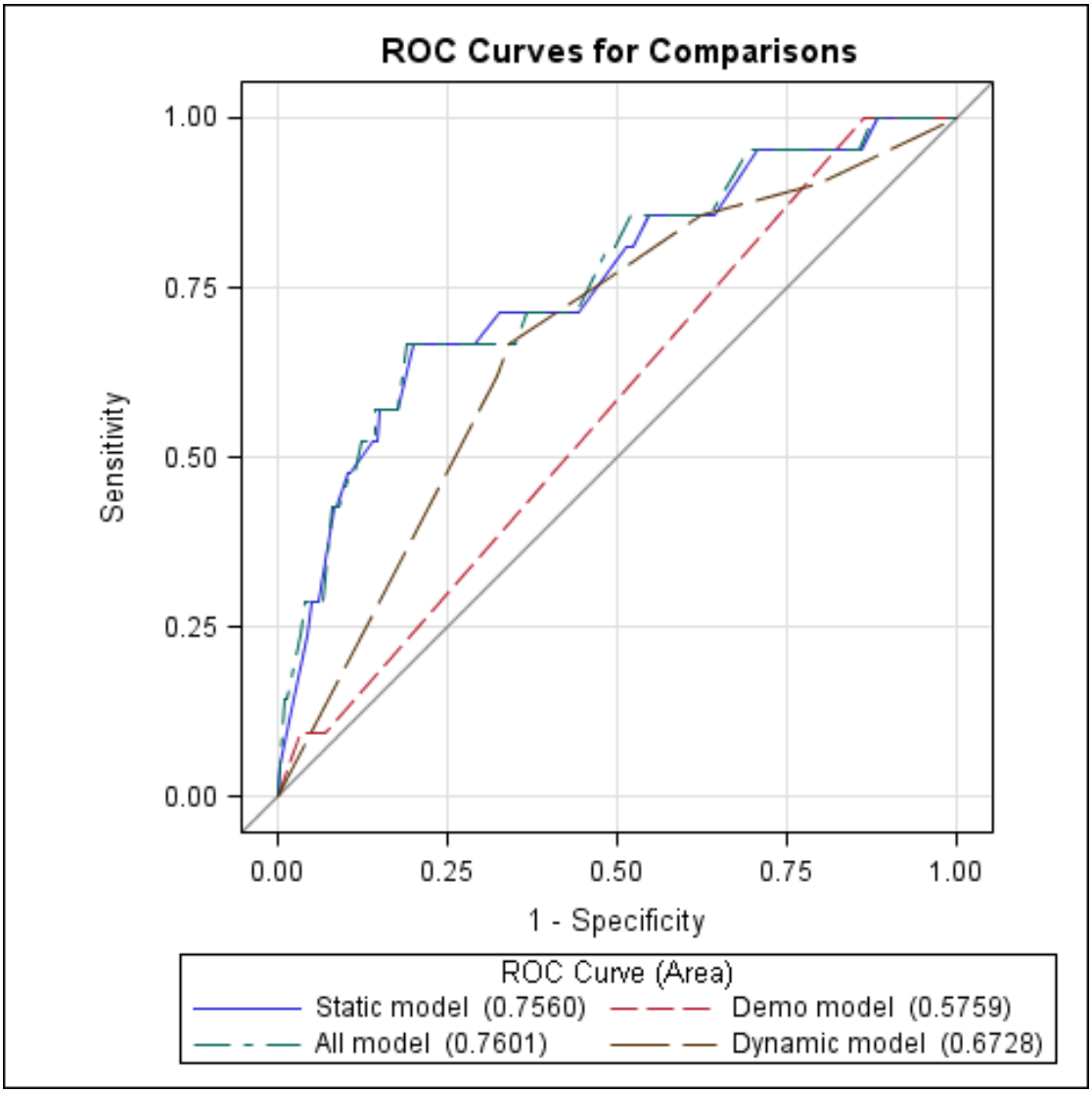


Figure 4.13 Receiver Operating Characteristic Curve (ROC) comparisons on life insurance data

4.4.7 Summary of the results

The case study illustrates how a SAAS is applied to South African life insurance data.

The organisational objectives were identified: to reduce payments to fraudulent claims and identify risk factors associated with fraudulent activity.

Data was made available for a sub-segment of the business, namely funeral products.

From the data, the risk factors were identified using techniques like anomaly detection, relational data and expert knowledge. An example of an anomaly is

the large claim amount indicator by loss class. An example of the use of relational data is the use of the suspicious agent indicator. An example of the use of expert knowledge is the time since inception to claim input. Based on the data availability and quality, a wide set of risk factors were made available for analysis.

The WOE measures and corresponding IVs were used to identify relevant risk factors (univariate analysis). To determine flexible split points for the WOE measures, a combination of statistical techniques and business considerations were used.

Fraud risk scorecards were constructed using WOE measures and backward logistic regression (multivariate analysis).

The scorecard using all inputs obtained an AUROC of 0.76 on the validation dataset. The AUROC value is reasonable.

Using business input, the cut-off across the posterior probability range may be determined based on the maximum profit or minimum cost. The cut-off may also be determined based on business input like the capacity of the investigations unit, for example.

4.5 CONCLUSION

In this chapter, as a preliminary investigation, the use of SMOTE was evaluated with promising results. Using a few inputs from the credit dataset (6 significant inputs), WOE and backward logistic regression, SMOTE improved the discriminatory performance of a scorecard from 0.745 to 0.753 value of the AUROC.

The study reported on two case studies using South African insurance datasets. In both case studies, the use of a SAAS was applied to construct fraud risk scorecards. WOE measures and IVs were used for scorecard construction at entity level. The risk factors were identified and evaluated using WOE measure and the IVs.

The literature and industry identified several risk factors that were thought to be useful for diagnostic fraud detection. Using South African specific insurance data, the empirical investigations confirmed the usefulness of

several of these risk factors to identify fraudulent activity, whilst others were found to be less predictive than expected.

Both case studies found strong association between specific regions and fraudulent activity. Several authors mention a strong association between geographic regions and fraudulent activity. The Italian Insurance Association reported that it noticed significant correlation between the region, claims frequency and percentage of fraudulent claims (Comité Européen des Assurances, 2007). The Insurance Fraud Bureau (2011) in the United Kingdom published a list of 'hot spots' where staged accidents occur more frequently. Musal (2010) used geographic analysis, peer-group analysis and regression to identify suspicious entities, namely service providers for medical insurance fraud detection.

Both case studies revealed that common demographic fraud risk indicators such as occupation, gender and age were statistically significant, but less significant than the static and dynamic indicators. The weak performance of these inputs may be attributed to data quality in the data samples that were available for empirical analysis.

In both case studies, the high frequency of claims by suspicious agents raise concern, as the exposure to fraud at agent level is much higher than the typical average claim amount associated with a fraudulent claim at policyholder level.

In the second case study, the statistical significance of the suspicious agent involvement indicator (strong IV of 0.364) makes the case of organised criminal activity stronger. The statistical significance of the suspicious agent indicator provides motivation for further research to explore the network of an entity in itself and the other entities in a suspicious entity's network. These relational risk factors may be better able to uncover networks of fraudulent activity or rings of organised criminals.

In both case studies, the time since inception to claim is a significant risk factor, which provides evidence to believe that fraudsters may start insurance policies with the intention to commit crime. It may provide an indication that the fraudulent activity is more often pre-meditated, rather than opportunistic.

In the first case study, textual concepts, namely roll-up terms, such as hi-jack and fire showed a statistically significant relationship with fraudulent activity.

The first case study on short-term insurance data (motor claims) obtained a reasonable AUROC value of 0.87 for the policyholder scorecard using a combination of demographical, static, dynamic and text data risk factors.

The agent scorecard obtained a good AUROC value of 0.86 on validation data.

The second case study on life insurance obtained an AUROC of 0.76 on the validation dataset.

The dynamic scorecard in the second case study performed worse than the static scorecard, which is in contrast with the results from the first case study. This deviance from the expected can be attributed to the nature of the insurance product – typically only one claim per policy (in the case of death), unless there are more than one entity or family member on the policy. This implies that the claims history is not as relevant as is the case for short-term insurance.

The risk factors that identify fraudulent activity include demographic, static claim, dynamic historic, textual and relational network information. The study found surprisingly good results with the application of a SAAS using basic demographic, static risk drivers and historical summaries.

In a production environment, the value of a fraud detection system is not easily quantified in monetary terms. The trade-off table and cost/savings matrix are useful tools to evaluate the impact of the diagnostic fraud detection scorecard. These tools are useful to evaluate the trade-off between the hit rate and false alarm rate of the scorecard and business constraints, like the capacity of the investigations department.

As more data becomes available for analysis, it would be possible to develop further scorecards for other entities such as service providers and assessors.

In addition, the study advocates the use of an entity-focused approach to fraud detection, given that fraudulent activity typically occurs at an entity or group of entities level. The human element of fraudulent activity should not be

disregarded. An entity level detection system would also support the work of specialist investigators.

Note that the intention of the empirical case studies was not to develop and report on the development of a fraud risk scorecard as such. Rather, the intention was to provide the reader with a better understanding of the scorecard construction process, as well as an understanding of the interpretation of WOE and how it applies to insurance data. In the case study, an entity focused approach works successfully.

The next chapter will conclude the study by providing a summary of the study, the key findings and subject areas proposed for future research.

CHAPTER 5

SUMMARY, FINDINGS AND CONCLUSION

'Now this is not the end. It is not even the beginning of the end.
But it is, perhaps, the end of the beginning'
Sir Winston Churchill

5.1 INTRODUCTION

The study presents itself within the context of statistical insurance fraud detection in South Africa. It outlines an approach, namely a SAAS, based on a probabilistic WOE framework that would be particularly useful for scorecard construction for fraud detection in the insurance industry in South Africa.

It attempts to bridge some of the divide between the complexity of analytical techniques and the interpretability and data challenges experienced when these techniques are utilised in the real world.

Despite several challenges, a statistical fraud detection approach would add value to the insurance industry in South Africa, both in the identification of fraud risk factors (descriptive analysis) and fraud detection (predictive analysis).

The first section of the current chapter provides a summary of the study by chapter. The following section discusses the key findings of the study. The section thereafter highlights the contributions of the study, followed by a section on possible topics for future research. The chapter concludes with closing remarks in the last section.

5.2 SUMMARY OF THE STUDY

In Chapter 1, the extent and context of insurance fraud were discussed. Current traditional and analytical approaches to fraud detection were discussed with examples from the literature. Technical research papers on the use of statistical techniques for fraud detection are bountiful, although little research exists specific to the insurance industry in South Africa. The insurance industry is aware of the magnitude of the insurance fraud problem in South Africa and started the South African Insurance Fraud Bureau in 2008.

The data collaboration platform should provide further opportunity for data-driven fraud detection.

In addition, in Chapter 1, several challenges with analytical fraud detection were discussed. These included the following:

- **Fraud detection is an n-class classification problem**

  The types of fraud are diverse and fraud schemes unique. The study made distinction between internal and external fraud and identified three types of external fraud from the literature: opportunistic, repeat and organised offences. The diversity in the types of fraud and the entities responsible complicates the use of a single scorecard or predictive model for supervised classification.

- **Suspicious activity are typically investigated after-the-fact, rather than pro-actively**

  The traditional reactive and inductive approach to detect fraud, leads to lengthy and costly investigations. As elucidated further in Chapter 2, the use of scorecards may be used as a pro-active and predictive tool for fraud detection. It would lessen the burden on specialist investigators by providing the specialist investigators with reason codes from the scorecard to reduce the audit cost.

- **A tension exists between the need to maximise profits and invest in anti-fraud measures**

  The study mentioned that insurance organisations are hesitant to invest in anti-fraud measures due to several reasons. In addition, the study identified a linear relationship between the complexity of the analytical techniques and the data requirements and amount of effort to support them.

- **A lack of resources exist in the special investigation teams**

  Owing to the large number of incoming claims, it is impossible for specialist investigators to screen each incoming claim manually. A data-driven semi-automated fraud detection approach, as outlined in the study, becomes imperative.

- **Known fraud cases form a skewed and small marginal class distribution**

  Known fraudulent cases are rare and amount to a very small proportion of all the claims received, which complicates the use of traditional supervised classification approaches. As discussed, to address this challenge, the class distribution can be changed using over sampling and under sampling techniques or by creating synthetic cases, such as SMOTE. Another approach is the use of misclassification costs. A final approach mentioned in Chapter 2 is the use of a classification algorithm that is insensitive to the class distribution in the training set. However, a common technique to find better split points is the use of traditional decision tree algorithms, which are not insensitive. The study explained that an alternative algorithm, such as HD hierarchical clustering, CC+ or HD decision trees, might find better split points that are not sensitive to class imbalance. More testing is required.

- **Fraudulent behaviour is dynamic**

  The study mentioned that fraudsters change their behaviour in order for their fraudulent activities to remain unnoticed. A SAAS should be able to adapt to the changing behaviour of fraudsters. The risk scorecards require model and input monitoring and regular revisions.

- **The value of fraud detection as a function of time**

  The study mentioned the importance of fast detection of fraudulent activity. For example, to develop a scorecard at claimant level, priority should be given to data available at the claims processing stage to ensure that the scorecard detects the fraudulent activity as soon as possible.

The study created a SAAS as outlined in Chapter 2, with the aim to address some of the challenges mentioned in Chapter 1. The SAAS provides a framework into which analytical fraud risk scorecards would operate and fit. A six-step methodology was suggested, which is not purely based on the input from quantitative analysis, but should take into account the expert knowledge of specialist investigators.

The six steps are summarised as follows:

1. **Identify organisational objectives**

   The study suggested that, together with the definition of the project parameters, that the cost and benefit of the SAAS should be determined in this step.

2. **Collect the data**

   The study recommended that the availability and quality of the data for fraud detection should be considered. The effort required to explore the data and prepare the data for analysis should not be underestimated.

3. **Evaluate segmentation**

   The study recommended that when scorecards are constructed, segmentation rules should be used to differentiate between the types of fraud and entities who commit fraud. Further segmentation may improve the performance of the fraud risk scorecards.

4. **Identify and evaluate the risk factors**

   The study evaluated the use of a WOE framework for risk factor evaluation for fraud detection. Both the relevancy and redundancy of the inputs should be considered. The WOE measures, IV and AUROC may be used to evaluate the relevancy, whilst Pearson correlation tests and variable clustering may be used to evaluate the redundancy of the inputs.

5. **Construct risk scorecards at entity levels**

   The study suggested that a staged implementation approach be followed for risk scorecard construction at entity level. The use of logistic regression and the WOE framework is recommended as it is transparent, robust and allow for an easy transformation into the scorecard format.

6. **Validate scorecards**

   The study recommended that validation statistics such as the the ROC Curve, AUROC, Cumulative Accuracy Profile (and corresponding AR), the Confusion Matrix, the F measure, the geometric mean and the cost matrix be used to evaluate the discriminatory power of the scorecards. It also

suggested that the misclassification costs be used to evaluate the performance of the scorecard, together with the cost matrix.

In addition, in Chapter 2, the functional form of the WOE measures and corresponding linear model were explained and illustrated with examples from the credit-scoring field. Several other benefits listed in Chapter 2 motivate the application of the WOE approach to fraud detection. These include that WOE measures are robust against missing values and outliers (as these are included and addressed in the discretisation process) and straightforward to interpret, which allows for an understandable and transparent risk assessment. The use of logistic regression also offers several benefits, one of which is that the effect of over sampling may be adjusted for, using the offset feature. In addition, the same offset feature may be used to include the misclassification costs in the model. The misclassification costs for fraud detection are typically unequal and should be considered.

Furthermore, in Chapter 2, the study explained that some authors found that popular binning algorithms, like Entropy-based decision trees, are not robust against unbalanced datasets. The study proposes the consideration of alternative binning algorithms, such as the CC+ measure, HD hierarchical clustering or HD decision tree. Further research is required to evaluate the usefulness of the HD decision tree to find better split points.

In the last section of Chapter 2, the use of several validation statistics was provided and explained. The validation statistics include the ROC Curve, AUROC, Cumulative Accuracy Profile (and corresponding AR), the Confusion Matrix, the F measure, the geometric mean, and the cost matrix.

The use of the WOE measures and risk factors provides a good starting point for data-driven fraud detection from which the types of information, in terms of risk factors and rules can evolve. The study mentioned that the monitoring of scorecard stability, input stability and population drift is critical for fraud detection. The use of a detection system will affect the behaviour of the fraudsters, called reactive population drift by Hand (2007). Perpretators will change their behaviour based on the detection system in place. The use of a SAAS, together with the WOE framework and a wide variety of techniques for

risk factor identification, would enable organisations to detect changes in fraudulent behaviour and update their fraud risk scorecards in a timely manner.

Chapter 3 elaborated on the use of data-driven techniques to identify more risk factors for fraud detection. Chapter 3 explained one popular approach: the use of anomaly detection. An anomaly is any deviation from the norm. Outlier detection, profiling and Benford's law may be used for anomaly detection. The second approach for risk factor identification, discussed in Chapter 3 was the use of text-mining algorithms on textual data fields. Text mining algorithms may identify risk factors located in the free-form text fields of organisations. Text mining algorithms may also be used to improve the data quality and usefulness of data. Thirdly, the use of relational diagnostics, identified by using social networks analysis, was explained.

Chapter 4 reported the findings of a preliminary and two empirical investigations.

The preliminary investigation evaluated the use of SMOTE to improve scorecard performance. Based on the credit dataset, it was found that the predictive performance of the scorecard improved slightly when using SMOTE. The AUROC of the scorecard improved from 0.745 to 0.753 by using SMOTE. However, it was decided not to include it in subsequent analysis as the actual distribution of the minority class was used and explored using the WOE. The actual distributions were used not only for prediction but also for exploration of the risk factors, whilst SMOTE add synthetic cases. However, SMOTE may add value when low numbers of known fraud cases are available for predictive analysis in a production environment.

The first insurance specific case study used short-term insurance data. Real-world data is not data-mining ready, so several steps were taken to improve the quality and usability of the data from the data warehouse. Data from several data marts were combined to form analytical base tables for scorecard development at entity level. Based on the availability of the data, the fraud detection scorecards were developed at the policyholder level and agent level. The final policyholder scorecard used a combination of demographical, static,

dynamic and text data, with an AUROC measure of 0.87 on the validation dataset. The final agent scorecard displayed an AUROC measure of 0.86 on the validation dataset.

The second insurance specific case study used life insurance data. Owing to a lack of data, the fraud detection scorecards were developed at claimant level only, although the agent code was included as a statistically significant risk factor. The final scorecard displayed an AUROC of 0.76 on the validation data.

5.3 KEY FINDINGS OF THE STUDY

The goals of the study and how it was achieved are summarised below:

The first goal of the study was to identify and define the challenges associated with an analytical fraud detection approach, applicable to the South African insurance industry.

It is evident from the literature that the magnitude of insurance fraud in South Africa is large. Surprisingly, very little research is available on statistical fraud detection within the context of the South African insurance industry.

The following challenges with an analytical fraud detection approach were identified and defined:

- Fraud detection is an n-class classification problem
- Suspicious activity are typically investigated after-the-fact, rather than pro-actively
- A tension exists between the need to maximise profits and invest in anti-fraud measures
- A lack of resources exist in the special investigation teams
- Known fraud cases form a skewed and small marginal class distribution
- Fraudulent behaviour is dynamic
- The value of fraud detection as a function of time

To address some of these challenges, a SAAS and probabilistic WOE framework is ideally suited for its task. The WOE measures allow insurance

organisations to explore and interpret risk factors and incorporate the combined effects of the risk factors using logistic regression. The scorecard format based on the transformation of the posterior probabilities into an additive scale allows for straightforward interpretation. During the claims screening process, the scorecard would allow the user to determine the reasons and context why entities or activities are flagged for fraud.

The second goal of the study was to create a fraud risk assessment system into which fraud risk scorecards would fit, to address specific challenges within the context of the South African insurance industry.

The study created a SAAS as a six-step approach which utilises quantitative analysis and the input from experts. The study defined WOE measures, corresponding IVs and logistic regression for scorecard construction. The study explained that fraud risk scorecards are robust against data issues like missing values, outliers and rare cases. In addition, fraud risk scorecards are easy to interpret and are therefore ideal as a decision support tool.

However, according to the literature, some popular binning algorithms that are used to find superior split points for scorecard construction, like Entropy-based decision trees, are not robust against unbalanced datasets. The study proposes the use of alternative binning algorithms, such as the CC+, hierarchical clustering algorithms and HD decision trees.

The third goal of the study was to apply analytical techniques, such as text mining, to uncover risk factors to improve the discriminatory performance of the fraud risk scorecards.

The study evaluated techniques, such as text mining, to identify risk factors to improve the performance of the fraud risk scorecards. Text mining, anomaly detection and network analysis proved itself useful in the identification of risk factors as discussed and applied empirically.

In both case studies, anomalies were detected by outlier indicators of the claim amount by loss class and, in the first case study, the frequency of claims.

In the first case study, using text mining algorithms, key concepts were extracted from the free-form text field, namely the claim narrative. Concepts

were included as risk factors in the form of roll-up terms. The roll-up terms representing hi-jack and fire demonstrated a statistically significant relationship with fraudulent activity. Furthermore, text mining algorithms were used to improve the quality of the data by performing a fuzzy merge on names from different data sources, using the generalised edit distance. In addition, text mining algorithms were used to extract the vehicle registration number from the claim narrative. Using the vehicle registration number, further summary information were obtained and used in subsequent analysis. The sum of claims using similar registration details demonstrated a statistically significant relationship with fraudulent activity.

The literature and industry identified several risk factors that proved successful to identify fraudulent activity in the past. Using South African insurance data, the empirical investigations confirmed the usefulness of several of these risk factors to identify fraudulent activity, whilst others were found to be less predictive than expected.

The risk factors that identify fraudulent activity include demographic, static claim information, dynamic historic information and relational network information.

In both case studies, demographic inputs like the geographic region and the occupation demonstrated a statistically significant relationship with fraudulent activity. These are in line with results from other authors (*cf.* Section 4.5). In the first case study, age and gender demonstrated a weaker statistically relationship with fraulent activity compared to static and dynamic inputs.

In both case studies, static inputs provided noticeable improvement of the ability of the scorecard to discriminate between fraudulent and non-fraudulent activity.

In the first case study, dynamic inputs provided further improvement of the discriminatory power of the fraud risk scorecard on policyholder level, although the same kind of lift was not observed in the second case study, which is attributed to the nature of the insurance product. In short-term insurance, the claims history provided valuable discriminatory improvement.

The performance of a SAAS was evaluated empirically using real-world insurance data with promising results. The study found surprisingly good results using basic demographic, static risk drivers and basic historical summaries (an AUROC of 0.87 observed using short-term insurance data and an AUROC of 0.76 observed using life insurance data).

Evaluate the application of fraud risk scorecards at entity level using South African insurance data.

The fourth goal of the study was to evaluate the application of a fraud risk assessment system at entity level using South African insurance data.

The study advocates the use of an entity-focused approach to fraud detection, given that fraudulent activity typically occurs at an entity or group of entities level. The human element of fraudulent activity should not be disregarded. The study developed fraud risk scorecards at entity level. Surprisingly, the entity level agent scorecard in the first case study displays comparative results (AUROC of 0.86). In the second case study, the suspicious agent indicator was significant (IV of 0.364). A limitation of the study was a lack of data to construct a fraud risk scorecard at agent level or at other entity levels in the second case study. As more data on the subject of other entities (*cf.* Figure 3.4) becomes available, more scorecards may be constructed.

Given the volumes of structured and unstructured data available in insurance organisations' data stores, it is possible to aid the fraud investigation process using supervised classification, like scorecard construction in the context of the South African insurance industry. The study found that the use of fraud risk scorecards for fraud detection is useful for exploratory and predictive analysis within the context of the South African insurance industry.

5.4 CONTRIBUTIONS OF THE STUDY

The study proposes the use of a SAAS and WOE measures for scorecard construction for diagnostic fraud detection, within the context of the South African insurance industry.

The study is novel in its proposal of scorecard construction for fraud detection within the context of the South African insurance industry. The use of scorecards for fraud detection will enable insurance organisations to identify

the statistically significant risk factors from a large set of potential risk factors, evaluate the relationship between the inputs and the outcome variable and control the process of the inclusion of risk factors within a SAAS. The study suggests that attention should be given to the use of binning algorithms during scorecard construction using unbalanced datasets. Insensitive measures, such as HD decision trees may be more suitable to determine split points, although the use of HD decision trees requires more testing.

The study recommends that careful consideration should be given to both the types of information as well as the types of statistical techniques in the design of a fraud detection system within the South African insurance industry. The types of information refer to the categories of input data available for analysis in the data warehouse, where consideration should be given to the quality and accessibility of the data. The types of statistical techniques refer to the constraints and assumptions of the underlying statistical techniques.

Today, the insurance industry and law enforcement agencies have access to information on a vast set of potentially predictive fraud risk factors. These may be demographical, contextual, behavioural or relational risk factors. These are gained from practical experience, documented in the literature, highlighted in news articles on fraud convictions and noted in research papers. Without statistical acumen and a centralised platform of data, organisations struggle to translate these risk factors into a fraud-risk assessment system with actionable outputs. The study explained the importance of a SAAS to identify where fraudulent activity exist and provide reasons for its classification. In addition, the uses of several data-driven techniques to identify predictive risk factors were evaluated. These include anomaly detection, text mining and social network analysis. All of these techniques proved valuable for risk factor identification in the empirical case studies.

The study is novel in its advocacy of the use of an entity-focused approach to fraud detection, given that fraudulent activity typically occurs at an entity or group of entities level. The human element of fraudulent activity should not be disregarded. An entity level detection system would also support the work of specialist investigators. The study identified the typical entities involved in an

insurance claim (*cf*. Figure 3.4). The identification of the suspicious entity is important, but so is the context of the fraudulent activity. Access to the behavioural history of suspicious entities and their networks would lead to shorter investigation periods, which in itself mean that potentially larger amounts may be recovered or the payment of fraudulent claims prevented.

5.5 FUTURE RESEARCH

There are several areas of opportunity for future research, which are listed and discussed below.

The use of other data sources for fraud risk factor identification

While insurance organisations have used credit bureau data in ratemaking and underwriting for several years (Wu & Guszczka, 2003), the relationship between credit data and fraud trends is under explored. Taking Cressey's fraud triangle (1972) into account, intuitively there is the notion that where there is more financial pressure, the motivation to commit fraud may increase. In line with the subject on the use of credit data for fraud detection and given the increase in fraudulent activity during the recent financial crisis, another topic of interest would be the study of the relationship between fraud trends and macro-economic data.

Given the increased availability of social media data (social media websites like Twitter, Facebook and the like), especially those of the unstructured kind, the use of textual and network data for fraud detection may become increasingly pertinent. Integrating social network data for fraud detection into the entity-level fraud risk scorecards may be useful to identify serious fraud activities, such as those committed by organised crime rings.

Determining optimal split points

According to the literature (Cieslak *et al.*, 2011), the use of HD decision trees would identify better split points than Entropy based decision trees when using unbalanced datasets. The use of these techniques, including the techniques used in the study, allows for improved positioning of the split points across the ranges of input values at a univariate level. Although some studies exist (Hand & Vinciotti, 2003), more can be done to identify the optimal number and positions of the split points and possibly extend the analysis to a multivariate

level. More research in this area would benefit both the credit scoring and analytical fraud detection subject areas.

Combination of several entity-level fraud risk scores

The use of network data was introduced in the study and remains an area where limited research is available on its practical use for fraud detection. Graph theory may be one of several approaches to enhance the entity-level fraud risk scores by including the scores of several connected entities to obtain an enhanced score for a particular entity. Alternatively, the scores at entity level may be used to calculate an overall score for a network. Again, for further research on fraud detection, specifically to identify fraud networks, the use of graph theory may show itself fruitful and interesting.

5.6 CONCLUSION

Insurance fraud remains a big challenge for the insurance industry, regulatory authorities and policyholders in South Africa. Evidence from the literature and the empirical case studies show the prevalence of insurance fraud. Insurance fraud is big business for opportunists, repeat offenders and organised criminals.

The volumes of structured and unstructured data within the data stores of organisations and within the public domain will continue to grow. Insurance organisations have volumes of data but lack a pro-active, transparent strategy and centralised platform for fraud detection. The popular saying of John Naisbitt holds true that organisations are drowning in information but starving for knowledge. Data quality remains a big challenge.

In addition, insurance organisations face several challenges when using a data-driven approach to fraud detection. The study outlines an approach within a framework, namely a SAAS, to utilise risk scorecards for fraud detection. The study also provides ways to transform unstructured text and relational network data into risk factors that can be used as inputs into the fraud risk scorecards.

The use of an analytical approach to fraud detection should form part of a holistic fraud management strategy, which include fraud prevention policies and procedures, internal controls and auditing. The use of a data-driven

SAAS will not solve all of the challenges related to insurance fraud; however, the use of such a system may lessen the burden on the specialist investigators and provide important insights for the timely detection of fraud.

Based on the promising results obtained in both empirical case studies, a combination of a quantitative (the use of data-driven risk factors) and qualitative analysis (evaluation of expert rules and tweaking of bins) within the SAAS seem appropriate. Moreover, in the case studies, the fraud risk scorecards display meaningful improvements when using the results from the text and network data analyses, and performing the analyses at entity level.

The fraud risk scorecards would enable an insurance organisation to recognise and detect fraudulent activity more accurately and rapidly, prioritise and improve the quality and quantity of their investigations, reduce their fraud expenditure, and uncover organised crime.

This study concludes with one final formula, obtained from Gracey *et al*. (2009):

The universal equation says:

$$\text{Incidence of fraud} = \frac{\text{the inclination + the opportunity}}{\text{the resistance}}$$

By increasing organisations' resistance against those who are inclined and have opportunity to commit insurance fraud, the incidences of insurance fraud must decline.

BIBLIOGRAPHY

ABRAHAMS, C. & ZHANG, M. 2009. Credit risk assessment: the new lending system for borrowers, lenders and investors. Hoboken, N.J.: Wiley. (Wiley and SAS business series.) 306 p.

ADAMS, R. 2010. Prevent, protect, pursue: a paradigm for fighting fraud. Computer fraud & security, 2010(7):5-11, July.

ALBRECHT, S.W., ALBRECHT, C., ALBRECHT, C.O. & ZIMBELMAN, M. 2009. Data-driven fraud detection. Fraud examination. 3rd ed. Mason, Oh.: South-Western. 663 p.

APPARAO, G., SINGH, A., RAO, G.S. & BHAVANI, B.L. 2009. Financial statement fraud detection by Data Mining. International journal of advanced networking and applications, 1(3):159-163.

ARTIS, M., AYUSO, M. & GUILLEN, M. 1999. Modelling different types of automobile insurance fraud behaviour in the Spanish market. Insurance: mathematics and economics, 24(1-2):67-68, 31 March.

BALDOCK, T. 1997. Insurance fraud. Canberra: Australian Institute of Criminology. (Trends and issues in crime and criminal justice, no 66.)

BARNES, P. & WEBB, J. 2003. Reducing an organization's susceptibility to occupational fraud: factors affecting its likelihood and size. London: BDE Global Business Risk and Security Management. http://www.bdeglobal.com/ Date of access: 21 August 2007.

BASEL COMMITTEE ON BANKING SUPERVISION. 2004. International convergence of capital measurement and capital standards: a revised framework. Basel: Basel Committee on Banking Supervision. 251 p.

BENFORD, F. 1938. The law of anomalous numbers. Proceedings of the American Philosophical Society, 78(4):551-572.

BERMUDEZ, L., PEREZ, J.M., AYUSO, M., GOMEZ, E. & VAZQUEZ, F.J. 2008. A Bayesian dichotomous model with asymmetric link for fraud in insurance. Insurance: mathematics and economics, 42(2):779-786.

BERRY, M.J.A. & LINOFF, G.S. 2000. Mastering data mining: the art and science of customer relationship management. 2nd ed. New York: Wiley Computer. 494 p.

BERZAL, F., CUBERO, J., MARIN, N. & SANCHEZ, D. 2004. Building multi-way decision trees with numerical attributes. Information sciences, 165:73-90.

BLONDEL, V.D., GUILLAUME, J.L., LAMBIOTTE, R. & LEFEBVRE, E. 2008. Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment, 10, October.

BOLTON, R.J. & HAND, D.J. 2002a. Statistical fraud detection: a review. Statistical science, 17(3):235-255.

BOLTON, R.J. & HAND, D.J. 2002b. Unsupervised profiling methods for fraud detection. Statistical science, 17(3):235-255.

BOLTON, R.J., HAND, D.J., PROVOST, F. & BREIMAN, L. 2002. Statistical fraud detection: a review. Statistical science, 17(3):235-254.

BONABEAU, E. 2002. Agent-based modeling: methods and techniques for simulating human systems. Proceedings of the National Academy of Sciences of the United States of America, 99(10):7280-7287.

BROCKETT, P.L., DERRIG, R.A., GOLDEN, L.L., LEVINE, A. & ALPERT, M. 2002. Fraud classification using principal component analysis of RIDITs. Journal of risk & insurance, 69:341-371.

CAHILL, M.H., LAMBERT, D., PINHEIRO, J.C. & SUN, D.X. 2002. Detecting fraud in the real world. (In Abello, J., Pardalos, P.M. & Resende, M.G.C., eds. Handbook of massive data sets. Dordrecht: Kluwer. p. 911-931.)

CHAN, P.K. & STOLFO, S.J. 1998. Toward scalable learning with non-uniform class and cost distributions: a case study in credit card fraud detection. (Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining. p. 164-168.)

CHAWLA, N.V., BOWYER, K.W., HALL L.O. & KEGELMEYER, W.P. 2002. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16(1):341-378, January.

CIESLAK, D.A. & CHAWLA, N.V. 2008. Learning decision trees for unbalanced data. (Proceedings of the 2008 European Conference on Machine Learning. Berlin: Springer. p. 241-256.)

CIESLAK, D.A., HOENS, T.R., CHAWLA, N.V. & KEGELMEYER, P. 2011. Hellinger Distance Decision Trees are Robust and Skew-Insensitive. The Journal of Data Mining and Knowledge Discovery. Preprint.

COALITION AGAINST INSURANCE FRAUD. 2011. Coalition against insurance fraud. http://www.insurance fraud.org Date of access: 28 April 2011.

COMITE EUROPEEN DES ASSURANCES. 2007. Comité Européen des Assurances. Annual Report 2007/2008. http://www.cea.eu/ Date of access: 28 April 2011.

CRESSEY, D.R. 1972. Criminal organization: its elementary forms. New York: Harper & Row. 127 p.

CROWN PROSECUTION SERVICE. 2006. Fraud Act. http://www.cps. gov.uk/ legal/d_to_g/fraud_act/ Date of access: 1 June 2009.

D.M. DISNEY & ASSOCIATES INC. 2010. D.M. Disney & Associates Inc. website. http://www.dm disney. com/insurance_fraud_indicators.htm Date of access: 10 July 2010.

DE KOKER, L. 2007. Financial crime in South Africa. Economic affairs, 27(1):34-38, March.

DE LA REY, T. 2007. Two statistical problems related to credit scoring. Potchefstroom: North-West University. Potchefstroom Campus. (Thesis - PhD. (Risk Management.))

DEBRECENY, R.S. & GRAY, G.L. 2010. Data mining journal entries for fraud detection: an exploratory study. International journal of accounting information systems, 11(3):157-181, September.

DELOITTE. 2011. Global insurance industry outlook. http://www.deloitte.com Date of access: 28 April 2011.

DOFF, R. 2007. Risk management for insurers: risk control, economic capital and solvency II. London: Risk Books. 204 p.

ENGELMANN, B., HAYDEN, E. & TASCHE D. 2003. Measuring the discriminative power of rating systems. Frankfurt am Main: Deutsche Bundesbank. (Series 2, Banking and financial supervision, no 01/2003.) 24 p.

FAWCETT, T. & PROVOST, F. 2002. Adaptive fraud detection. Data mining and knowledge discovery, 1:291-316.

FELDMAN, R. & SANGER, J. 2006. Text mining handbook: advanced approaches in analyzing unstructured data. Cambridge: Cambridge University Press. 410 p.

FICO. 2010. FICO Falcon fraud manager technology highlights. http://www.fico.com/ en/products/dmapps/pages/fico-falcon-fraud-manager.aspx Date of access: 10 July 2010.

FINNY, H.C. & LESIEUR, H.R. 1982. A contingency theory of organizational crime. Research in the sociology of organizations, 1:255-299.

FRANCIS, L.A. 2006. Taming text: an introduction to text mining. Casual Actuarial Society Forum: 51-88, Winter.

GALIMI, J. & EARLEY, A. 2005. Gartner: deploy detection technologies to cut insurance fraud and abuse. http://www.gartner.com/DisplayDocument? id=476971& ref=g_sitelink. Date of access: 9 January 2006.

GEONAMES. 2011. Geonames. http://geonames.org/. Date of access: 5 February 2011.

GOOD, I.J. 1960. Weights of evidence, corroboration, explanatory power, information and the utility of experiments. Journal of the Royal Statistical Society. Series B (Methodological), 22(2):319-333.

GRACEY, B., COLLINS, J., JONES R., PLUMB, S., TILLEY H., WELFARE, R. & WILLIAMS, D. 2009. Fraud investigation: a claims handler's guide. (A practical guide to the key issues and current law.)

GREENACRE, M.J. 1993. Correspondence analysis in practice. San Diego, Calif.: Academic Press. 195 p.

HAND, D.J. 2007. Statistical techniques for fraud detection, prevention, and evaluation. London: Imperial College. http://videolectures.net/mmdss07_hand_stf/ Date of access: 15 November 2010.

HAND, D.J. & ADAMS, N.M. 2000. Defining attributes for scorecard construction in credit scoring. Journal of applied statistics, 27(5):527-540.

HAND, D.J. & VINCIOTTI, V. 2003. Scorecard construction with unbalanced class sizes. Journal of the Iranian Statistical Society, 2(2):189-205.

HARTDEGEN, P. 2009. Ensure all is right with insurance. http://www.property24.com/articles/ensure-all-is-right-with-insurance/10748 Date of access: 15 November 2010.

HILAS, C.S. & MASTOROCOSTAS, P. 2008. An application of supervised and unsupervised learning approaches to telecommunications fraud detection. Journal knowledge-based systems, 21(7):721-726, October.

HILL, S., PROVOST, F. & VOLINSKY, C. 2007. Learning and inference in massive social networks. (The 5th International Workshop on Mining and Learning with Graphs, August 2007.)

HILL, T.P. 1995. A statistical derivation of the significant-digit law. Statistical science, 10:354-363.

HOSSACK, I.B., POLLARD, J.H. & ZEHNWIRTH, B. 1999. Introductory statistics with applications in general insurance. New York: Cambridge University Press. 282 p.

HUANG, S.M., YEN, D.C., LUEN-WEI, S.M., YANG, L.W. & HUA, J.S. 2008. An investigation of Zipf's Law for fraud detection. Decision support systems, 46(1):70-83.

INSURANCE FRAUD BUREAU. 2011. Insurance Fraud Bureau. http://www.insurancefraud bureau.org/ Date of access: 28 April 2011.

JANS, M., LYBAERT, N. & VANHOOF, K. 2007. Data mining for fraud detection: toward an improvement on internal control systems. (Proceedings of the 30th Annual Congress European Accounting Association (EAA2007). Lisbon, Portugal.)

JOU, S. & HEBENTON, B. 2007. Insurance fraud in Taiwan: reflections on regulatory effort and criminological complexity. International journal of the sociology of law, 35:127-142.

JUSZCZAK, P. & ADAMS, N.M. 2008. Off-the-peg and bespoke classifiers for fraud detection. Computational statistics & data analysis, 52(9):4521-4532, May.

KLEINBAUM, D.G. 1994. Logistic regression: a self-learning text. New York: Springer. 282 p.

KROLL AMERICA. 2009. Global Fraud Report. Annual edition, 2009/2010. Kroll America. http://www.krollfraudsolutions.com/media Date of access: 28 April 2011.

KUBAT, M., HOLTE, R.C. & MATWIN, S. 1998. Machine learning for the detection of oil spills in satellite radar images. Machine learning - Special issue on applications of machine learning and the knowledge discovery process, 30(2-3):195-216, Feb/March.

LASALLE, R.E. 2007. Effects of the fraud triangle on students risk assessments. Journal of accounting education, 25(1-2):74-87.

LECKEY, A. 2009. Fraud schemes tend to follow economic trends. News article. http://www.chicagotribune.com/business/yourmoney/sns-yourmoney-0517scams,0, 6511545.story Date of access: 1 June 2009.

LEWIS, D.D. & GALE, W. 1994. A sequential algorithm for training text classifiers. (Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. New York: Springer.)

LIU, W., CHAWLA, S., CIESLAK, D.A. & CHAWLA, N.V. 2010. A robust decision tree algorithm for imbalanced data sets. (Proceedings of the SIAM International Conference on Data Mining, SDM 2010, April 29 - May 1, 2010, Columbus, Ohio, USA. p. 766-777.)

LRP PUBLICATIONS. 2009. Insurance data shows jump in fraudulent claims linked to recession. News article. Risk & insurance. (LRP publications.) http://www.riskandinsurance. com/story.jsp?storyId=212728307 Date of access: 1 June 2009.

MAJOR, J.A. & RIEDINGER, D.R. 2002. A hybrid knowledge/statistical-based system for the detection of fraud. Journal of risk and insurance, 69(3):309-324.

MCDONALD, R.A., STURGESS, M., SMITH, K., HAWKINS, M.S. & HUANG, E.X. 2011. Non-linearity of scorecard log-odds. International Journal of Forecasting. Article in Press. Corrected Proof.

MENA, J. 2003. Investigative data mining for security and criminal detection. Burlington, Md.: Elsevier Science. 452 p.

MORLEY, N.J., BALL, L.J. & ORMEROD, T.C. 2006. How the detection of insurance fraud succeeds and fails. Psychology, crime & law, 12(2):163-180, April

MUSAL, R.M. 2010. Two models to investigate medicare fraud within unsupervised databases. Expert systems with applications, 37:8628-8633.

NAKAJIMA, C. 2007. Editorial issue: Issues on fighting financial crime. Economic affairs, 27(1):2-5. March.

NAVARRO, G. 2001. A guided tour to approximate string matching. ACM computing surveys, 33(1):31-88, March.

NEWCOMB, S. 1881. Note on the frequency of use of the different digits in natural numbers. American journal of mathematics, 4(1):39-40.

NIGRINI, M.J. 1999. Adding value with digital analysis. Internal auditor, 56(1):21-23.

PADMAJA, T.M., DHULIPALLA, N., BAPI, R.S. & KRISHNA, P.R. 2004. Unbalanced data classification using extreme outlier elimination and sampling techniques for fraud detection. (Transactions of the 15th International Conference on Advanced Computing and Communications. p. 511-515.)

PHUA, C., ALAHAKOON, D. & LEE, V. 2004. Minority report in fraud detection: classification of skewed data. ACM SIGKDD explorations newsletter - Special issue on learning from imbalanced datasets, 6(1):50-59, June.

PHUA, C., LEE, C., SMITH, K. & GAYLER, R. 2005. A comprehensive survey of data mining-based fraud detection research. Artificial intelligence review. 14 p..

PORTER, D. 2007. The evolution of fraud intelligence. (In Rao, H.R., Gupta, M. & Upadhyaya, S., eds. Managing information assurance in financial services. Hershey, Pa.: IGI Publication. p. 261-283.)

PRWEB. 2009. Alarming increase in insurance fraud. News article. http://www. prweb.com/releases/2009/05/prweb2399334.htm Date of access: 25 May 2009.

ROBB, D. 2004. Text mining tools take on unstructured data. Computer world, 1 June.

ROBINSON, P. 2007. The FSA's perspective on insurance fraud. Speech. http://www.fsa.gov.uk/pages/Library/Communication/Speeches/2007/0926_pr.shtml Date of access: 29 September 2009.

SARMA, K.S. 2005. Combining decision trees with regression in predictive modeling with SAS Enterprise Miner. (SUGI 30. Proceedings. Philadelphia, Pennsylvania, April 10-13, 2005. Paper 074-30.)

SAS INSTITUTE INC. 2009. Getting started with SAS(R) Enterprise Miner 6.1. Cary, NC: SAS Institute.

SAS INSTITUTE INC. 2010. Predictive modelling using logistic regression. Cary, NC: SAS Institute.

SCHILLER, J. 2007. The impact of insurance fraud detection systems. Journal of risk and insurance, 73(3):421-438.

SIDDIQI, N. 2006. Credit risk scorecards: developing and implementing intelligent credit scoring. Hoboken, N.J.: Wiley. (Wiley and SAS business series)

SOUTH AFRICAN INSURANCE CRIME BUREAU. 2008. South African Insurance Crime Bureau website. http://www.saicb.co. za/news/3-south-african-insurance-crime-bureau-launched.html Date of access: 10 January 2009.

STEFANOWSKI, J. & WILK, S. 2008. Selective pre-processing of imbalanced data for improving classification performance. Lecture notes in computer science, 5182:283-292.

STOKES, G. 2010. The war on insurance fraud. News article. http://www.fanews.co.za/article.asp?Front_Page_Features; 25, Featured_Story; 1147, The_war_on_insurance_fraud;8261 Date of access: 6 January 2011.

SUBELJ, L., FURLAN, S. & BAJEC, M. 2010. An expert system for detecting automobile insurance fraud using social net work analysis. Expert systems with applications, 38:1039-1052.

SUTHERLAND, E.H., CRESSEY, D.R. & LUCKENBILL, F. 1992. Principles of criminology. 11th ed. Lanham, Md.: General Hall. 696 p.

TALEB, N.N. 2007. The Black Swan: The Impact of the Highly Improbable. New York: Random House Group. 366 p.

THOMAS, L.C., EDELMAN, D.B. & CROOK, J.N. 2002. Credit scoring and its applications. Philadelphia, Pa.: Society for Industrial and Applied Mathematics. (SIAM monographs on mathematical modeling and computation.) 248 p.

VAN DER MERWE, C.G. & DU PLESSIS, J.E. 2004. An introduction to South African law. The Hague: Kluwer Law International. 483 p.

VAN ZYL, H. 2010. South African Insurance Crime Bureau. Presentation. South African Underwriter Manager Association Conference. www.sauma.org. Date of access: 28 April 2011.

VIAENE, S., DEDENE, G. & DERRIG, R.A. 2005. Auto claim fraud detection using Bayesian learning neural networks. Expert systems with applications: an international journal, 29(3):653-666, October.

VIAENE, S., DERRIG, R.A. & DEDENE, G. 2004a. A case study of applying boosting naive bayes to claim fraud diagnosis. IEEE transactions on knowledge and data engineering, 16(5):612-620, May.

VIAENE, S., DERRIG, R.A. & DEDENE, G. 2004b. Cost-sensitive learning and decision making for Massachusetts PIP Claim Fraud Data. International journal of intelligent systems, 19:1197-1215.

VIAENE, S., DERRIG, R.A., BAESENS, B. & DEDENE, G. 2002. A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection. Journal of risk and insurance, 69(3):373-422.

VIANE, S., VAN GHEEL, D., AYUSO, M. & GUILLEN, M. 2004. Cost-sensitive design of claim fraud screens. (Industrial Conference on Data Mining, 2004. p. 78-87.)

WEISS, G.M. 2002. Mining with rarity: a unifying framework. ACM SIGKDD explorations newsletter - Special issue on learning from imbalanced datasets, 6(1):7-19.

WESTON, D., HAND, D.J., ADAMS, N.M., WHITROW, C. & JUSZCZAK, P. 2008. Plastic card fraud detection using peer group analysis. Advances in data analysis and classification, 2(1):45-62.

WHEELER, R. & AITKEN, S. 2000. Multiple algorithms for fraud detection. Knowledge-based systems, 13(2/3):93-99, April.

WILLIAMS G. 1998. Evolutionary hot spots data mining: an architecture for exploring for interesting discoveries. (PAKDD '99. Proceedings of the Third Pacific-Asia Conference on Methodologies for Knowledge Discovery and Data Mining. London: Springer.)

WILLIAMS, G.J. & HUANG, Z. 1997. Mining the knowledge mine: the hot spots methodology for mining large real world databases. (AI '97. Proceedings of the 10th Australian Joint Conference on Artificial Intelligence. Advanced Topics in Artificial Intelligence.)

WOODFIELD, T.J. 2004. Mining textual data using SAS Text Miner for SAS9. Course Notes. Cary, NC: SAS Institute.

WOODFIELD, T.J. 2005. Predicting workers' compensation insurance fraud using SAS Enterprise Miner 5.1 and SAS Text Miner. (SAS Institute white paper (071-30).)

WU, C.S.P. & GUSZCZA, J.C. 2003. Does credit score really explain insurance losses? Multivariate analysis from a data mining point of view. http://casualtyactuaries.com/pubs/forum/03wforum/03wf113.pdf. Date of access: 10 July 2010.

YAN, J., GUSZCZA, J., FLYNN, M. & WU, C.S.P. 2009. Applications of the offset in property-casualty predictive modeling. Casualty Actuarial Society E-Forum: 366-385, Winter.

YANG, W.S. & HWANG, S.Y. 2006. A process-mining framework for the detection of healthcare fraud and abuse. Expert systems with applications, 31(1):56-68, July.

ADDENDUM A

TECHNICAL DETAILS

1.1 Credit Data

A general credit scoring dataset from SAS, called accepts.sas7bdat, was used to explain analytical concepts. The dataset consists of 3000 loan applications with a binary outcome variable, default event (one if loan defaulted and zero otherwise). There are 18 numerical inputs and 8 categorical inputs available on the dataset.

1.2 Insurance Data

In the first case study, data from the data warehouse of a major short-term insurance organisation in South Africa was used, combined with a flat file from the insurance organisation's forensics department.

In the second case study, data from the data warehouse of a life insurance organisation in South Africa was used. The dataset contained a fraud indicator.

1.3 Software used

The SAS programming language was used for data preparation and exploration. The SAS programming language and SAS Enterprise Miner were used for model development. WEKA software was used to apply SMOTE.

1.4 Ethical Considerations

The datasets were anonymised before it was used in the empirical investigations.

1.5 Limitations of the study

- The study assumed that the data received from the insurance organisation was clean and accurate.
- The actual implementation of a SAAS resided outside the scope of the study.

ADDENDUM B

KEY THEORETICAL CONCEPTS

- **Fraud**

Any unlawful act or omission under which a misrepresentation is made with the intention to defraud which causes actual prejudice or which is potentially prejudicial to another, whether or not there is personal benefit to the perpetrator For example, when a policyholder colludes with an agent to claim for losses not really incurred.

- **Weights of Evidence (WOE)**

I.J. Good (1960) introduced the Weights of Evidence as a metric of the explanatory power of an attribute and defined it as follows:

$$W(H:E) = \log \frac{P(E|H)}{P(E|\overline{H})}.$$

where the purpose is to know the influence of independent variable E on the dependent outcome H.

- **Unbalanced Dataset**

An unbalanced dataset has unequal frequencies of the dependent outcome variable (also known as a skewed dataset). For example, known fraudulent claims can be as little as 1 percent of all claims received at an insurance organisation.

- **Robustness**

Robustness is a characteristic describing a model's, test's or system's ability to effectively perform while its variables or assumptions are altered. A robust concept can operate without failure under a variety of conditions.

- **Parsimonious model**

The simplest plausible model with the fewest possible number of inputs.

- **Text Mining**

The process that uses a set of algorithms to convert unstructured text into structured data objects.

- **Social Networks Analysis**

A study of social relationships in terms of vertices (nodes) and edges (ties), where the vertices generally represent persons (related to graph theory).

- **Credit Scoring**

A set of statistical techniques used to determine whether to extend credit to a borrower.

- **Numerical Input**

A numeric input is a continuous univariate classifier, also known as an independent integer variable. For example, the verifiable income of a policyholder is a numerical input.

- **Categorical Input**

A categorical input is a nominal univariate classifier, also known as an independent class variable. For example, the loss class category of a claim is a categorical input.

- **Event Rate (ER)**

The event rate is defined as the number of positive instances (where the outcome variable is equal to 1) divided by the number of all instances.

- **Class Confidence+ (CC+)**

The Class Confidence positive measure is defined as the number of positive instances (in an attribute) divided by the total number of positive instances (in the sample).

- **Hellinger Distance (HD)**

The Hellinger distance measures the divergence between two probability measures independent of the prior probabilities.

ADDENDUM C

EXAMPLES OF DIAGNOSTIC FRAUD INDICATORS

1. Demographic data

Diagnostic	Description	Reference
Age	Age of entity	Mena, 2003
Gender	Gender of entity	Mena, 2003
Region	Geographic region of residence	Mena, 2003
Occupation	Occupation of entity	Mena, 2003

2. Static data

Diagnostic	Description	Reference
High premium v. low legitimate income	High premium payment compared to verifiable legitimate income.	Robinson, 2007
No concern about costs	Lack of concern by the policyholder over charges or costs for early redemption.	Robinson, 2007
High commissions	Unusually high commissions paid to intermediary (agent).	Robinson, 2007
Far from agent	Insured live and work far from agent.	D.M. Disney & Associates, 2010

3. Dynamic data

Diagnostic	Description	Reference
Repeated beneficiary changes	Repeated and unexplained change of beneficiary.	Robinson, 2007
Previous suspicious claims	Previous redupiated claims based on suspicious information provided.	D.M. Disney & Associates, 2010
Policy change	Recent increase of policy limits	D.M. Disney & Associates, 2010

ADDENDUM D

PRESENTATION AT THE CREDIT SCORING AND CREDIT CONTROL CONFERENCE IN EDINBURGH, UNITED KINGDOM ON THE 28TH OF AUGUST 2009

Scorecard design for fraud detection (with text mining, predictive modelling and social network theory)

Terisa Roberts
Philip Pretorius

28 August 2009

NORTH-WEST UNIVERSITY
YUNIBESITI YA BOKONE-BOPHIRIMA
NOORDWES-UNIVERSITEIT

State of the Global Insurance Industry

- Global recession
- Acquisitions and Mergers
- Stronger competition, especially from non-traditional insurance companies
- Tighter regulation

"Fraudulent claims have doubled in the first three months of 2009" - Allianz Insurance, United Kingdom

State of South African Insurance Industry

- Emerging market
- Cost of insurance more expensive in relation with people's income
- South African insurance industry has tendency to follow international trends

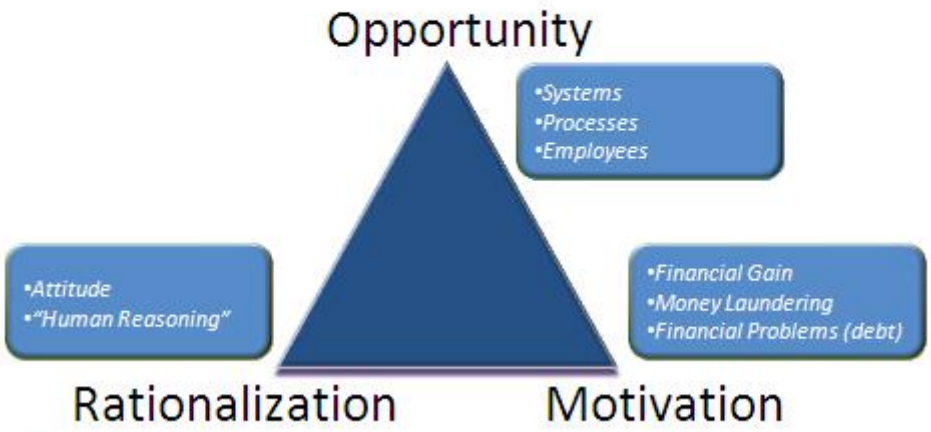
Insurance Fraud in South Africa

- 5%-10% of claims said to be fraudulent*
- Costing insurance industry R2 billion per year
- Set up of the South African Insurance Crime Bureau in 2008

"Between 8% and 35% of short-term insurance claims paid out to policyholders annually are fraudulent." – Insurance Companies, South Africa

*South African Insurance Association, 2008

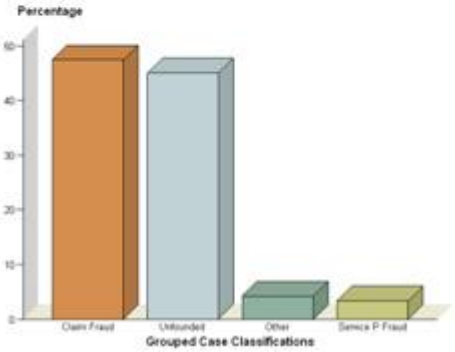
Fraud Triangle*

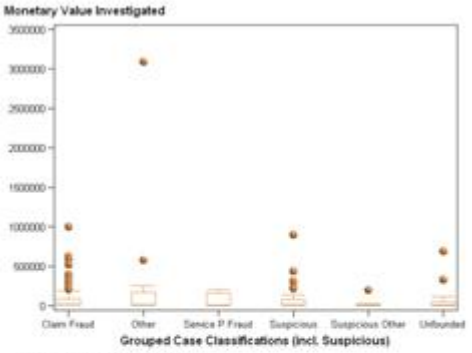


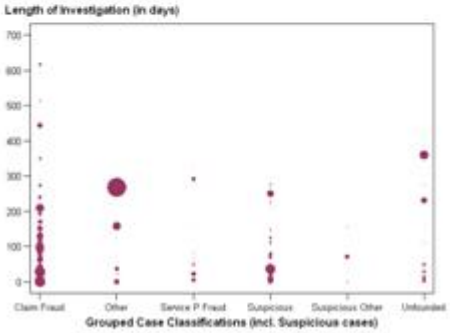
*Dr Donald R. Cressey, 1950

Investigated Cases

- Large proportion unfounded
- Mean(investigation period) = 80 days
- Other = internal fraud etc.







Challenges in Fraud Management

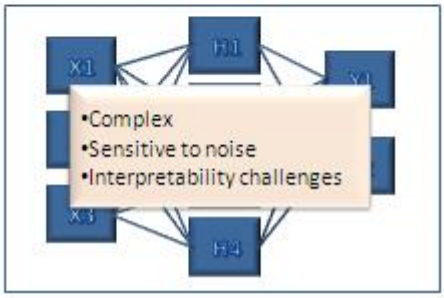
- Fraud detection reactive, rather than pro-active
- Diagnostic indicators commonly used – but not tested
- Special Investigations Unit – limited resources, fraud management spreadsheets
- "Feedback loop" not complete
- Infrequent event data, "tip of the iceberg"
- Detection techniques used with varying degrees of success
  - Redundant complexity (out-dated rules)
  - Disparate systems and utilization
  - Deteriorating performance over time

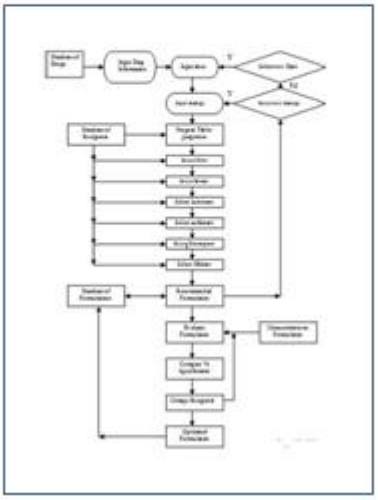
Fraud Detection Techniques

- Diagnostic Fraud Indicators
- 3rd party data searching
- Anomaly Detection
- Profiling
- Trained/Untrained Classifications
- Artificial Intelligence
- Text Mining
- Social Network Theory

Artificial Intelligence

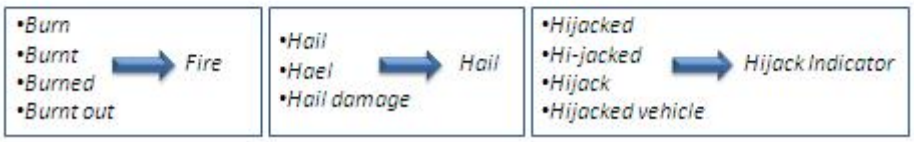
- Machine Learning
- Neural Networks
- Expert Systems
- Benford's Law



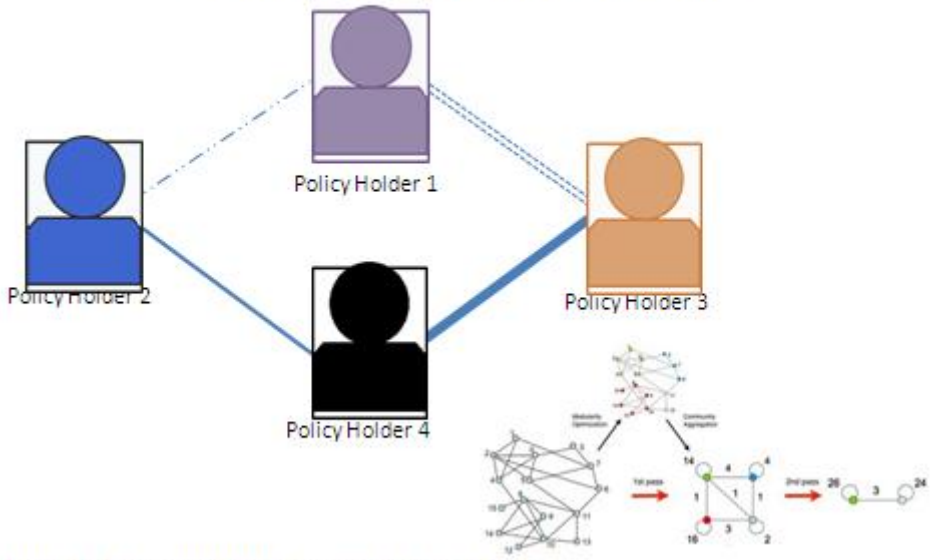


Text Mining

- Natural Language Processing
- Semantics – meaning of words
- Syntax – structural relationship between words
- Text Parsing
- Dimension Reduction Techniques



Social Network Theory



*Fast unfolding of communities in large networks, 2008

Key Properties* of a Suspicious Activity Assessment System

- Accurate
- Fast
- Cost-effective
- Flexible
- Consistent
- Reliable
- Easy to interpret
- Adaptive

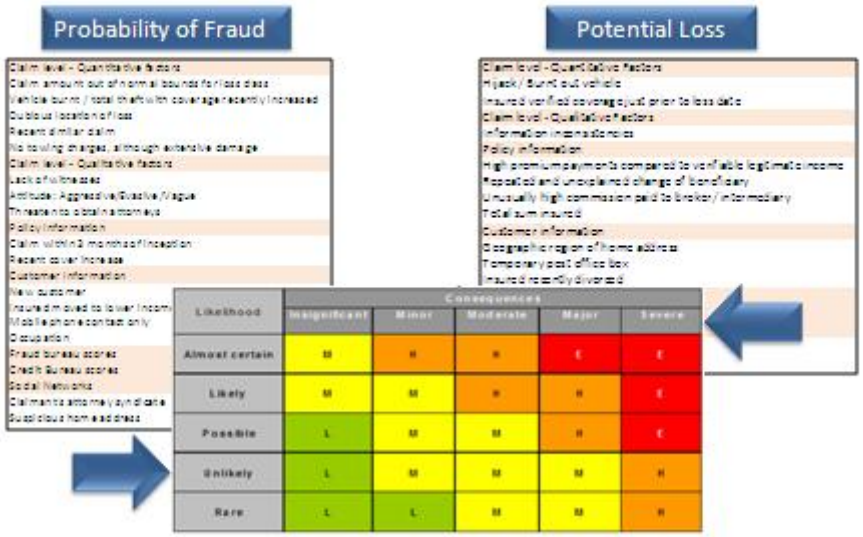
*Clark Abrahams, Credit Risk, 2008

The Case for a Suspicious Activity Assessment System

- "We are drowning in information, but starving for knowledge"*
- Combined qualitative and quantitative assessment
- Pro-active and Transparent

*Malik, 2006

(Hypothetical) Fraud Risk Scorecards



Conclusion

- Fraud remains a big challenge
- A pro-active and accurate suspicious activity assessment system should enable insurance companies to
  - Prioritise and improve quality and quantity of investigations
  - Reduce fraud expenditure
  - Uncover organised crime
- By utilising volumes of internal, external, structured and unstructured data
- Whilst maintaining an easy to implement and easy to interpret design

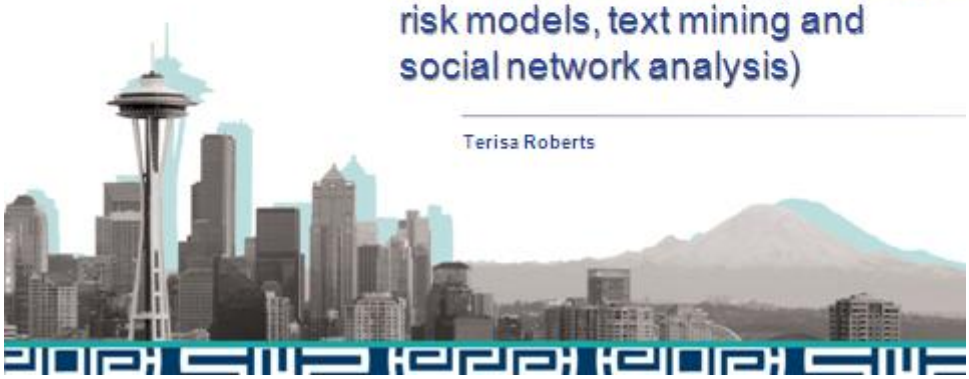
ADDENDUM E

PRESENTATION AT SAS GLOBAL FORUM IN SEATTLE, UNITED STATES OF AMERICA, ON THE 13TH OF APRIL 2010

SAS GLOBAL FORUM

Improving the defense lines ... The future of fraud detection in the insurance industry (with fraud risk models, text mining and social network analysis)

Terisa Roberts



SAS GLOBAL FORUM 2010

State of the global insurance industry

- Recession

"Fraudulent claims have doubled in the first three months of 2009" - Allianz Insurance, United Kingdom

State of the global insurance industry

- Acquisition and mergers
- Stronger competition
- Tighter regulation

Insurance fraud in South Africa

- Emerging market
- Cost of insurance more expensive in relation to people's income
- 5 - 10% of claims said to be fraudulent*

"Between 10% and 35% of short-term insurance claims paid out to policyholders are fraudulent."
– Insurance Companies, South Africa

*South African Insurance Association, 2008

Challenges in fraud management

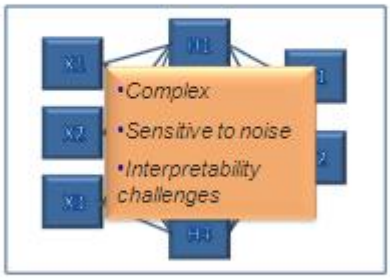
- Reactive detection techniques, rather than pro-active
- Infrequent event data, "tip of the iceberg"
- Limited resources in Special Investigations Unit
- "Feedback loop" not complete
- Diagnostic indicators used, BUT not tested

Traditional fraud detection techniques

- Whistleblower hotlines
- Internal audit procedures
- Watch lists
- Diagnostic fraud indicators
- Profiling
- Anomaly detection

Advanced analytical fraud detection techniques

- Predictive Modeling
  - Decision Trees
  - Neural Networks
  - Regression Analysis
- Expert Systems
- Text Mining
- Social Network Analysis

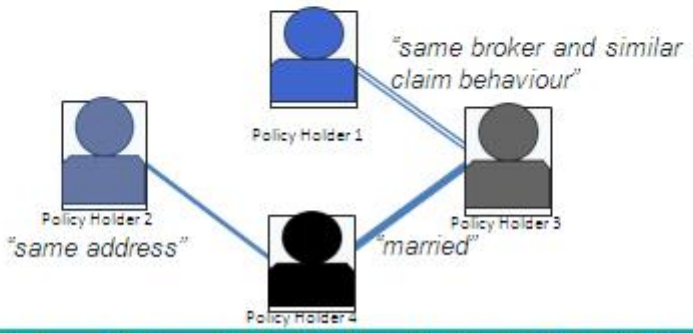


Text Mining

- Natural Language Processing
- Semantics – meaning of words
- Syntax – structural relationship between words
- Text Parsing
- Dimension Reduction Techniques: Roll up terms

Social Networks Analysis

- From Data – Links – (Community Detection) - Networks – Scores and Summaries
- Scores and summaries included in scorecards



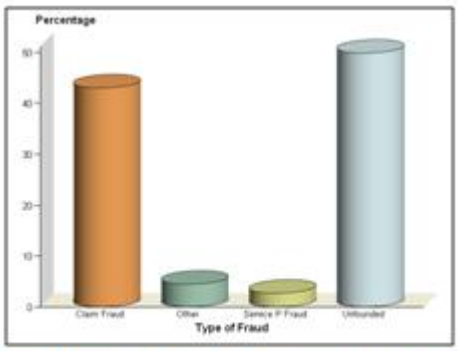
The case for a suspicious activity assessment system

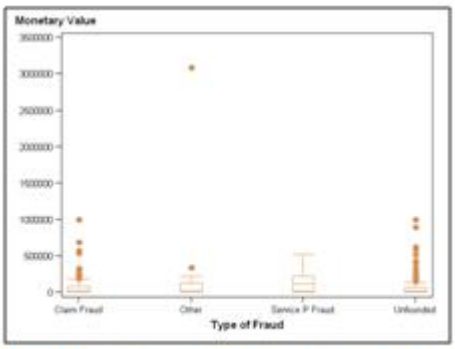
- Fast
- Accurate
- Cost-effective
- Consistent
- Flexible
- Easy to interpret
- Adaptive

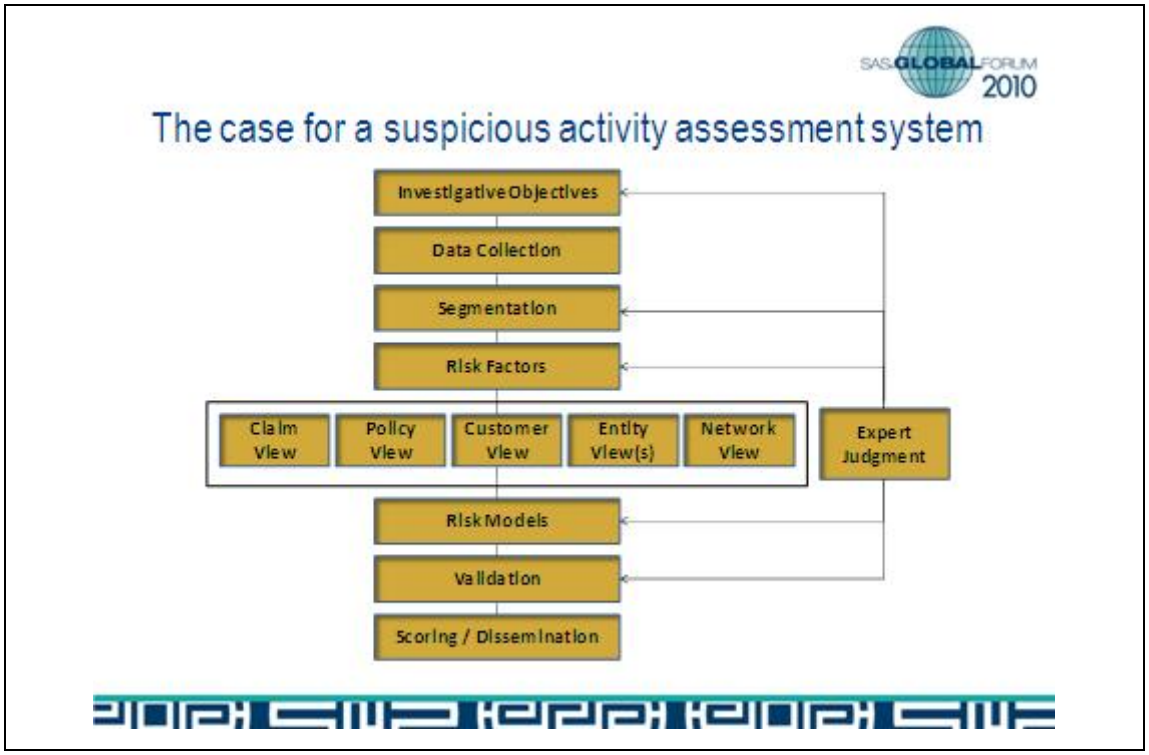
*Clark Abrahams, Credit Risk, 2008

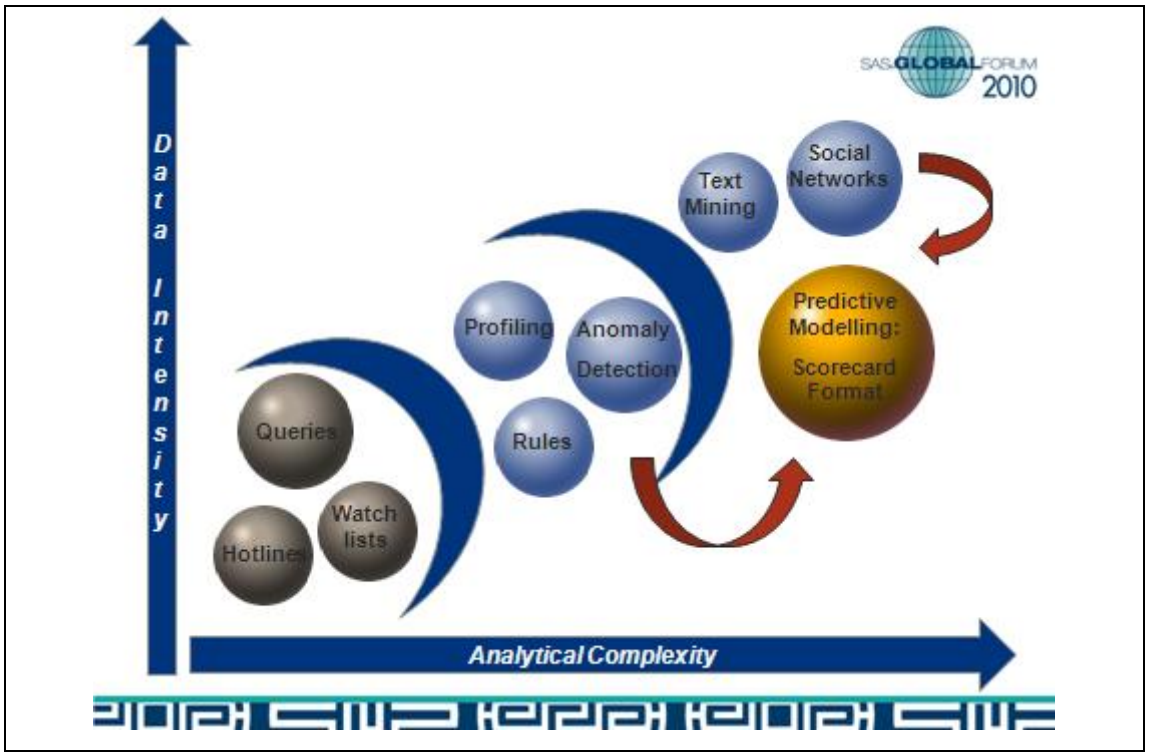
Case study : Insurance company data (fraud cases)

- Large proportion: Unfounded
- Average investigation: 80 days






SAS GLOBAL FORUM 2010
The case for a suspicious activity assessment system
Investigative Objectives
Data Collection
Segmentation
Risk Factors
Claim View
Policy View
Customer View
Entity View(s)
Network View
Expert Judgment
Risk Models
Validation
Scoring / Dissemination


SAS GLOBAL FORUM 2010
Data Intensity
Analytical Complexity
Text Mining
Social Networks
Predictive Modelling: Scorecard Format
Profiling
Anomaly Detection
Rules
Queries
Watch lists
Hotlines

Fraud Risk Models

- Scorecard format
  - WOE measures & Regression

Output Variables

Variable	Gini Statistic	Information Value	Level for Interactive	Exported Role
customer_riskfactor2	42.234	0.722	INTERVAL	Input
industry_class	36.806	0.61	ORDINAL	Input
customer_riskfactor1	41.315	0.53	INTERVAL	Input
claim_riskfactor1	32.192	0.449	INTERVAL	Input
claim_riskfactor2	30.585	0.341	INTERVAL	Input
network_riskfactor1	20.689	0.286	INTERVAL	Input
riskfactor9	22.171	0.237	ORDINAL	Input
pol_riskfactor1	26.054	0.225	INTERVAL	Input
riskfactor10	20.167	0.19	INTERVAL	Input
internal_rating	18.497	0.189	ORDINAL	Input
country_group	14.948	0.186	ORDINAL	Input
region_ind	14.468	0.09	NOMINAL	Rejected
customer_riskfactor10	10.042	0.082	ORDINAL	Rejected
riskfactor8	9.708	0.059	INTERVAL	Rejected
network_riskfactor2	10.939	0.045	INTERVAL	Rejected
riskfactor24	8.161	0.037	ORDINAL	Rejected
claim_riskfactor4	6.681	0.025	INTERVAL	Rejected

Scorecard

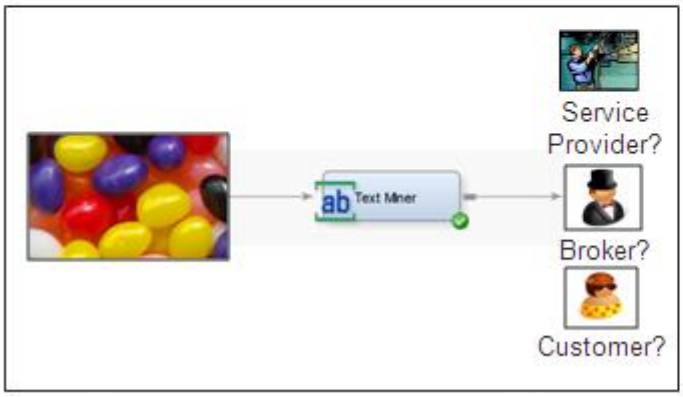
		Scorecard Points
claim_riskfactor1	claim_riskfactor1< -0.01	33
	-0.01<= claim_riskfactor1< 1.19	29
	1.19<= claim_riskfactor1	26
	Missing	13
claim_riskfactor2	claim_riskfactor2< 4.28	1
	4.28<= claim_riskfactor2< 6.88	19
	6.88<= claim_riskfactor2< 8.18, Missing	17
	8.18<= claim_riskfactor2< 8.88	61
	8.88<= claim_riskfactor2	37
country_group	1, 2	439
	3, 4	-11
	5, 7, 9, Missing	-34

Segmentation

- Wide spectrum of fraud

Segmentation

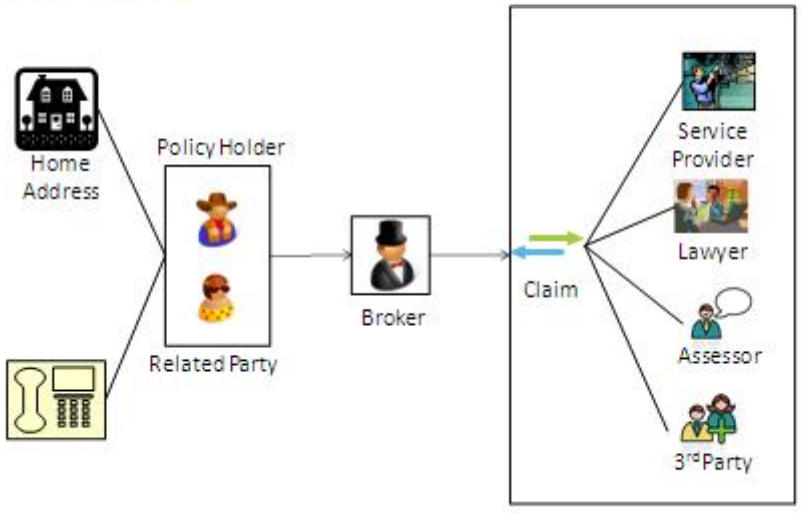
- Which entities are involved?



Terms

Term	Attribute
+ claim	Alpha
+ vehicle	Alpha
+ client	Alpha
+ assessor	Alpha
+ invoice	Alpha
fraud	Alpha
+ repair	Alpha
+ report	Alpha
policy	Alpha
+ damage	Alpha
+ investigation	Alpha
vat	Alpha
+ inflate	Alpha
+ number	Alpha
+ pay	Alpha
+ windscreen	Alpha
auto	Alpha
benefit	Alpha
sae	Alpha
+ panelbeaters	Alpha
+ recover	Alpha
broker	Alpha
company	Alpha

Entity Model

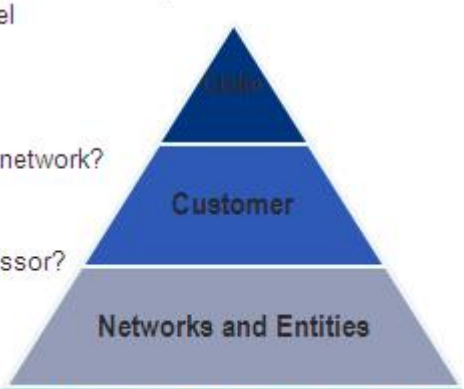


Risk Factors

- Diagnostic fraud indicators
  - Red flag rules based on industry knowledge
  - Judgmental indicators (i.e. Aggressive / Evasive attitude)
- Predictive variables
  - Available input variables
  - Claims, policy and customer level histories
  - Outlier analysis
- Novel data sources
  - Credit Bureau Information

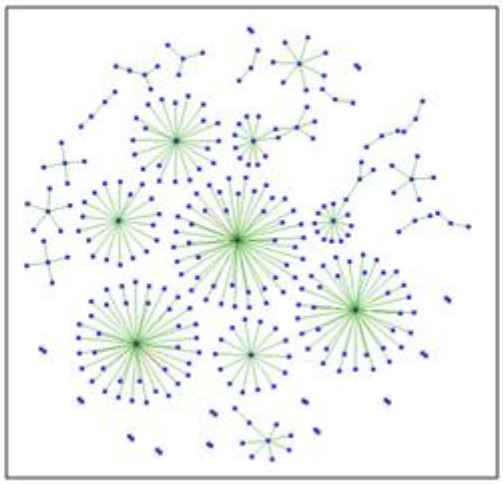
Views

- Claim / Policy / Customer
  - Traditionally, fraud models were developed at claim/policy/customer level
- Network
  - Suspicious network?
  - High volume of claims on network?
- Entity / Entities
  - Suspicious broker or assessor?



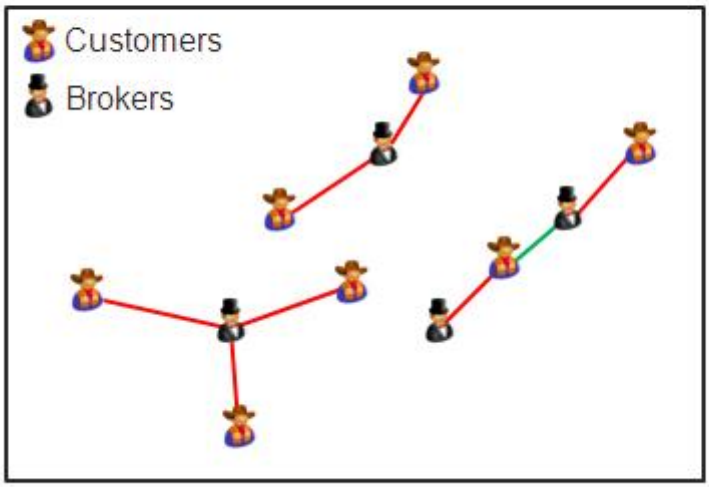
SAS GLOBAL FORUM 2010

Sub-Network of customers and brokers (by town)



SAS GLOBAL FORUM 2010

Deeper investigation shows suspicious relationships…



(Hypothetical) Fraud Risk Scorecards

Probability of Fraud

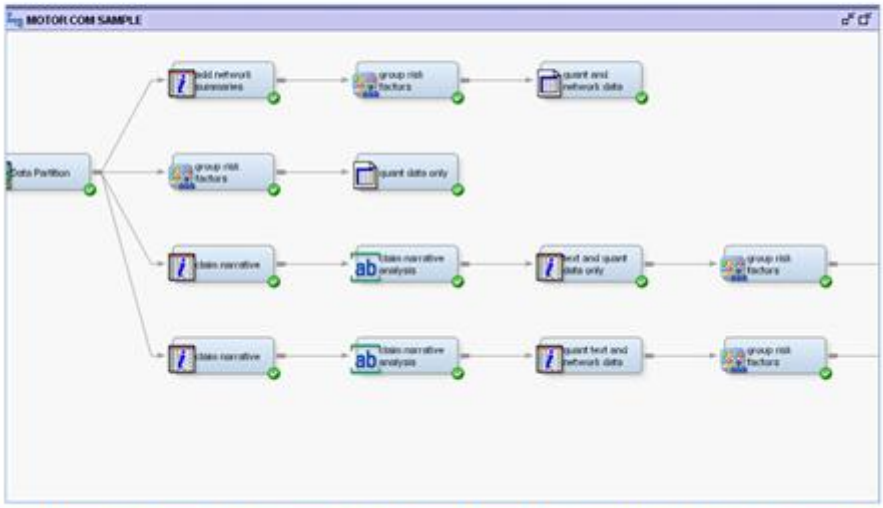
- Claim View
  - Quantitative Factors
    - Amount outside normal bounds for loss class
    - Vehicle Burnt
    - Dubious location of loss
    - *Recent similar claim
    - No towing charge although extensive damage
  - Qualitative Factors
    - Lack of witnesses
    - Aggressive / Evasive Attitude
    - Threaten to obtain attorneys
- Policy View
  - Claim within 3 months of inception
  - *Recent cover increase
- Customer View
  - Insured recently moved to low or income location
  - Mobile phone contact only
  - High risk occupation class
- Fraud Bureau Scores
- Credit Bureau Scores
- Social Network View
  - Suspicious attorney - involvement in syndicate
  - Suspicious home address

Potential Loss

- Claim View
  - Quantitative Factors
    - Hijack / Burnt out vehicle
  - Qualitative Factors
    - Discrepancies
- Policy View
  - High premium payments compared to verifiable income
  - *Repeated change of beneficiary
  - High total sum insured
- Customer View
  - Temporary post office box
  - *Recently divorced
- Fraud Bureau Scores
- Credit Bureau Scores
- Social Network View

Likelihood	Consequence				
	Insignificant	Minor	Moderate	Major	Severe
Almost Certain	M	H	H	E	E
Likely	M	M	H	H	E
Possible	L	M	M	H	E
Unlikely	L	M	M	M	H
Rare	L	L	M	M	H

Fraud Risk Models in SAS Enterprise Miner



Conclusion

- Fraud remains a big challenge
- A pro-active and accurate suspicious activity assessment system should enable insurance companies to
  - Prioritise and improve quality and quantity of investigations
  - Reduce fraud expenditure
  - Uncover organized crime
- Whilst maintaining an easy to implement and easy to interpret design

ADDENDUM F

PAPER SUBMITTED IN CONJUNCTION WITH PRESENTATION AT SAS GLOBAL FORUM IN SEATTLE, WASHINGTON ON THE 13TH OF APRIL 2010

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
Prof. Philip D. Pretorius

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.

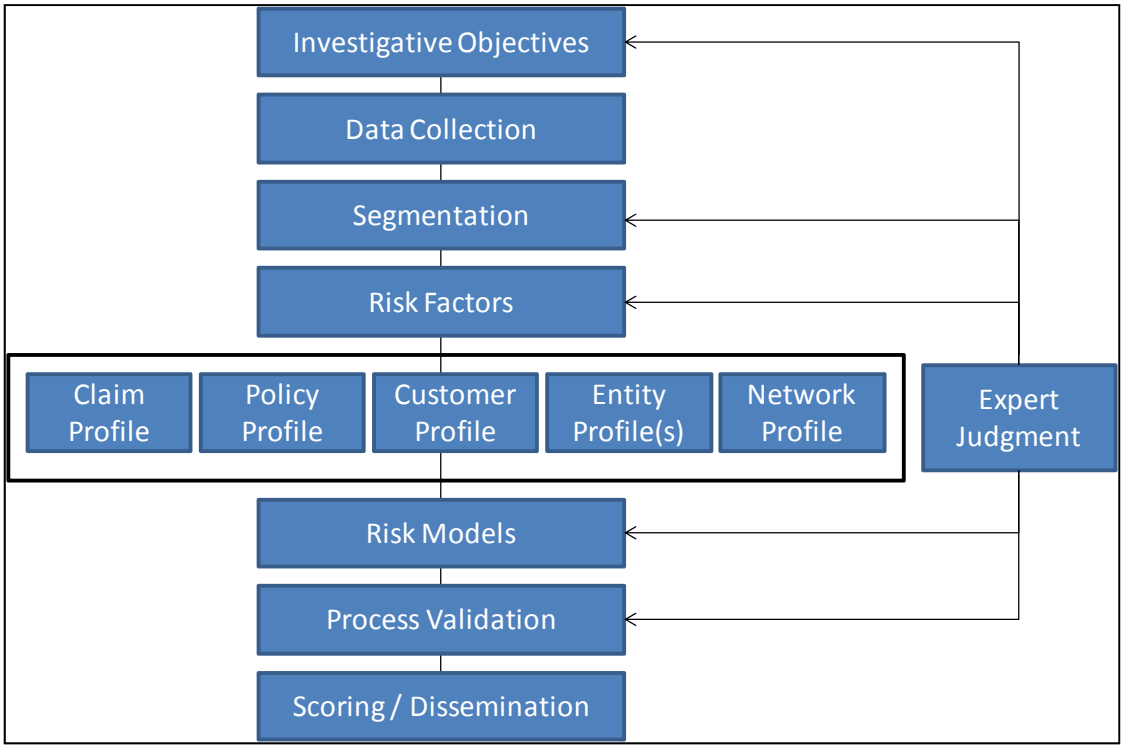


Figure 1. Suspicious Activity Assessment System: Analytical Process

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.

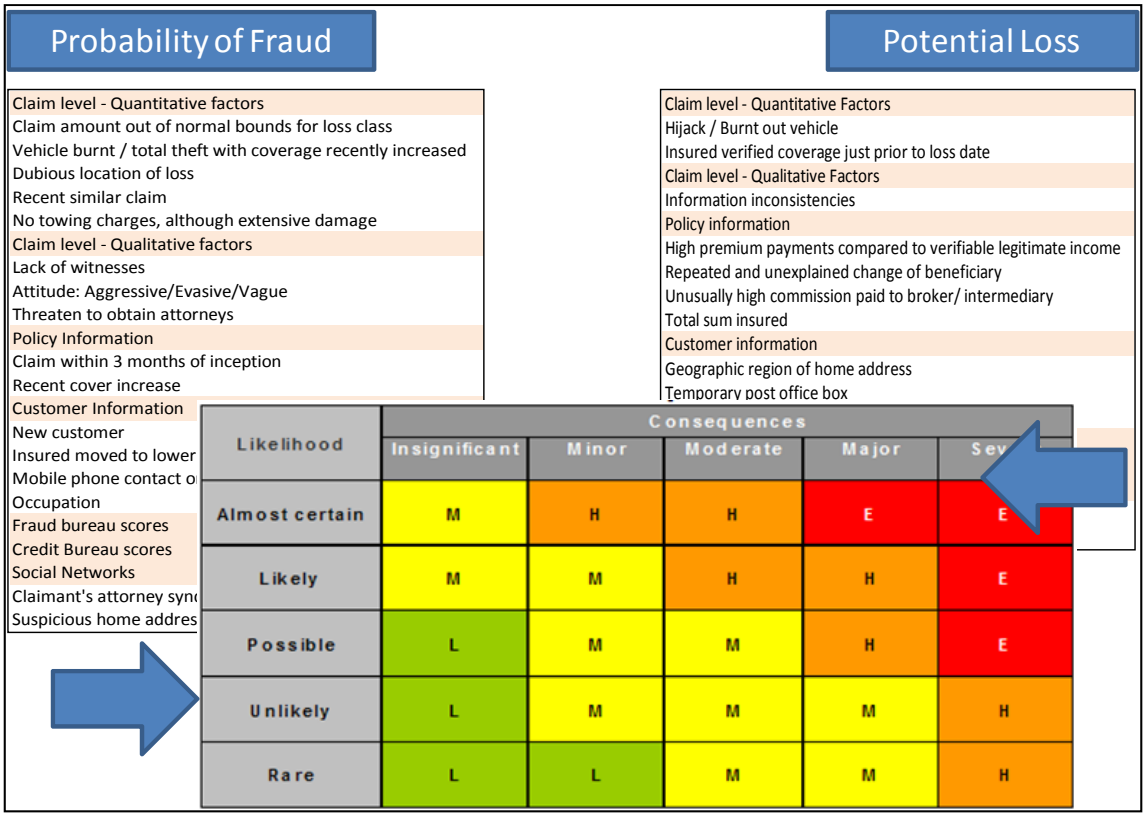


Figure 2. Hypothetical Example of Fraud Risk Scorecards

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, SAS® Social Network Analysis software provides a network visualization interface to present a complete picture of entities, their characteristics, and networks.

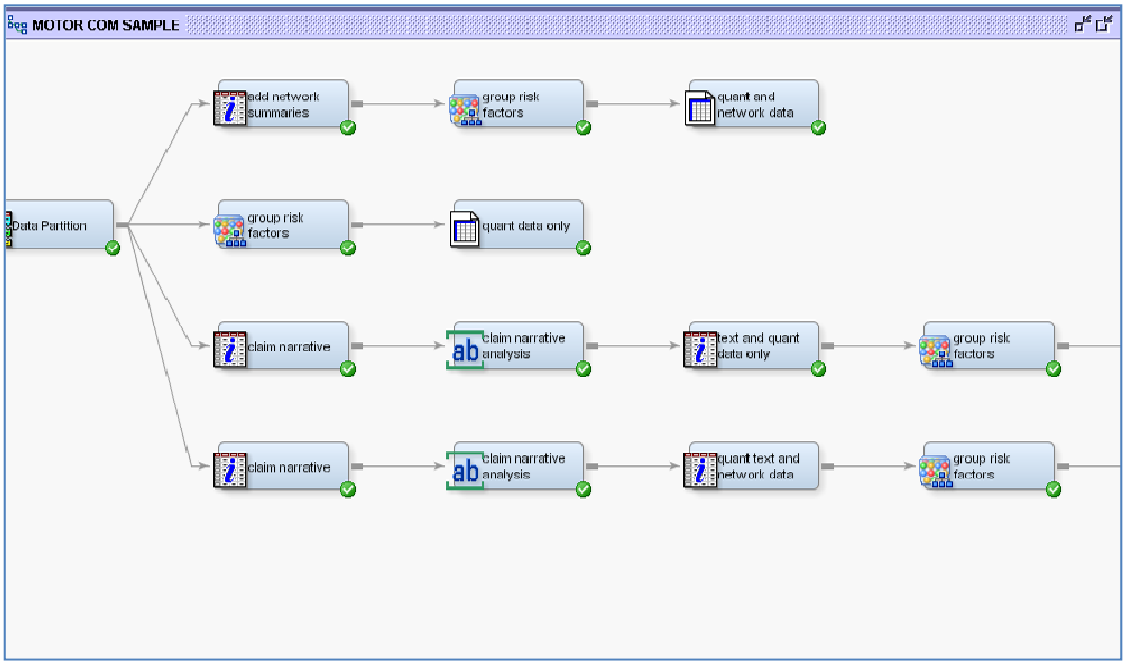


Figure 3. Fraud Risk Scorecards (with Text Data and Network Summaries) in SAS Enterprise Miner

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.