

INTELLIGENT INSURANCE CLAIMS FRAUD PREDICTION USING MACHINE
LEARNING

by

DAVID LEICESTER KENYON

Submitted in fulfilment of the requirements for the degree

Master of Science (Computer Science)

in the

FACULTY OF ENGINEERING, BUILT ENVIRONMENT AND
INFORMATION TECHNOLOGY

at the

UNIVERSITY OF PRETORIA

PRETORIA, SOUTH AFRICA

SUPERVISOR: Prof. J.H.P. Eloff

JUNE 2018

© University of Pretoria, 2018.

All rights reserved.

Abstract

Insurance fraud costs South Africa (and the global insurance industry) billions of Rands. Insurance claims fraud, which involves over-inflating claim amounts or fabricating a loss to result in a claim settlement, makes up a substantial portion of this cost. It would therefore be beneficial to the insurance industry to have a way of intelligently identifying insurance claims fraud. Current strategies focus on identifying fraud *after* the fact through methods such as auditing. These methods can be enhanced by predicting whether claims are fraudulent *before* they get paid, instead of after payment has already been made.

Techniques in the fields of data science and machine learning can be used to intelligently predict insurance claims fraud, based on existing data. Because insurers have large sets of data, it is suggested that Big Data be factored in when predicting insurance claims fraud. However, new and proposed privacy legislation requires data scientists to be mindful and consider privacy when mining users' personal information.

The current research addresses the problems of insurance fraud, data bloat and information privacy by proposing a framework, model and architecture. The proposed framework contains the processes necessary to intelligently predict insurance claims fraud. The model that is suggested can be used to predict insurance claims fraud. The architecture shows software and hardware components that can be used to create a prototype. The research as a whole discusses this prototype, how it was developed, and how it was tested.

Acknowledgements

I would like to thank the following people for their contribution to this research:

- Prof. J.H.P Eloff for his infinite wisdom and guidance throughout my research. His knowledge of the field of computer science and his continuing support and direction made this research possible.
- Fulcrum for their financial backing, infrastructure, study time and support. In particular, I would like to thank Paul Schreuder for always being understanding and allowing me the much needed time to perform this research. I would also like to thank Magesh Naidoo for her continued support, help and motivation.
- Chris, for always being there. Your insight, patience and kindness was a strong motivator throughout this study and I will forever be indebted to you.
- Lastly, I would like to thank my parents Carol and Colin for their encouragement and support during this time.

Contents

Abstract	iii
Acknowledgments	iv
List of Tables	1
List of Figures	2
1 Introduction	1
1.1 Purpose Statement	4
1.2 Problem Statement	4
1.3 Research Questions	5
1.4 Scope of the Study	6
1.5 Research Methodology Followed	7
1.5.1 Conducting a Literature Review	7
1.5.2 Developing a Research Use Case	7
1.5.3 Designing and Developing a Solution	7
1.5.4 Validating the solution	7
1.6 Terminology	7
1.6.1 Cyber Security	8
1.6.2 Intelligent Fraud Prediction	8
1.6.3 Insurance Claims Fraud	8
1.6.4 Big Data Science	9
1.6.5 Privacy	9
1.6.6 Solution	10
1.7 Layout of this Dissertation	10
2 The Insurance Industry	13
2.1 Issues in the Insurance Industry with regard to Data	13
2.1.1 Strategies	14
2.1.2 Operations	14
2.1.3 Regulations	15
2.1.4 Insurance Modernisation	18
2.2 Data in the Insurance Industry	20
2.2.1 Data in the Insurance Industry regarding Data Quality Tools	21
2.2.2 Data in the Insurance Industry regarding Data Standards	21
2.2.3 Data in the Insurance Industry regarding Insurance Claims	22

2.3	The Use of Data to Solve the Issues in the Insurance Industry	26
2.3.1	Strategies and Data	26
2.3.2	Operations and Data	26
2.3.3	Regulations and Data	27
2.3.4	Insurance Modernisation and Data	27
2.3.5	Comparison of Issues with Possible Solutions	27
2.4	Discussion	28
3	Fraud and Fraud Prediction in The Insurance Industry	29
3.1	Fraud	30
3.1.1	Definition: Insurance Fraud	30
3.1.2	Types of Insurance Fraud	30
3.1.3	Insurance Claims Fraud	31
3.1.4	Claims Fraud's Occurrence in Insurance Processes	34
3.2	Traditional Fraud Detection Techniques	40
3.2.1	Fraud Auditing	41
3.2.2	Forensic Accounting	41
3.2.3	Whistle-blowing	42
3.3	Intelligent Fraud Prediction	42
3.3.1	Definition: Intelligent Fraud Prediction	43
3.3.2	Intelligent Fraud Prediction: Financial Systems	44
3.3.3	Intelligent Fraud Prediction: Insurance Claims Systems	47
3.4	Discussion	51
4	Big Data Science	53
4.1	Big Data	53
4.1.1	Definition: Big Data	54
4.1.2	Variety	54
4.1.3	Velocity	55
4.1.4	Volume	55
4.1.5	Advanced Computing Requirements for Big Data	56
4.1.6	Hadoop Distributions	58
4.1.7	Hadoop Sub-projects	59
4.2	Data Science and Advanced Analytics	60
4.2.1	Definition: Data Science	60
4.2.2	Machine Learning	63
4.3	Data Science Platforms	68
4.3.1	KNIME	69
4.3.2	R	69
4.3.3	RapidMiner	69
4.4	Data Science: Privacy Consideration	70

4.5	Discussion	71
5	Proposed Framework for Intelligent Insurance Claims Fraud Prediction	73
5.1	Requirements	73
5.1.1	Business Requirements	74
5.1.2	Stakeholder Requirements	74
5.1.3	Transition Requirements	74
5.1.4	Functional Requirements	74
5.1.5	Non-functional Requirements	75
5.2	Framework	75
5.2.1	Services	76
5.2.2	Interfaces	77
5.2.3	Rules	77
5.2.4	High-level Component Diagram Showing the Fraudulent Claims Prediction Process	78
5.3	High-level Use Case Diagram Showing the Fraudulent Claims Prediction Process	79
5.3.1	Actors	79
5.3.2	Use Cases	80
5.4	Discussion	80
6	Detailed Design	83
6.1	Data Preparation	84
6.1.1	Data Exploration	85
6.1.2	Data Pre-processing	85
6.2	Model Creation	88
6.2.1	Model Training	89
6.2.2	Accuracy Testing	90
6.3	Knowledge Application	91
6.3.1	Creation of XML Rules based on Apriori Association Rules	91
6.3.2	Addition of Logistic Regression Model to Systems	91
6.3.3	Automatic Recommendations	91
6.3.4	Derivation of Formula	94
6.3.5	Insurance Claims Fraud Prediction Model (ICFPM)	100
6.4	Model Maintenance	103
6.5	Summary	103
6.6	Discussion	104
7	Architecture	105
7.1	Operational Requirements	106
7.2	Technical Requirements	107

7.3	Quality Requirements	109
7.3.1	Manageability	109
7.3.2	Security	109
7.3.3	Performance	110
7.4	Layers	110
7.5	Structure of the System	111
7.5.1	Software View	111
7.5.2	Hardware View	113
7.6	Discussion	114
8	Constructing and Validating The Prototype	115
8.0.1	Construction of the Prototype	115
8.0.2	Validation of the Prototype	116
8.1	Constructing the Prototype	116
8.1.1	Generation of the Prototype: Insurance Claims Management System	117
8.1.2	Generation of the Test Data	120
8.1.3	Cleaning the Test Data	122
8.1.4	Filtering the Test Data	123
8.1.5	Transforming the Test Data	124
8.1.6	Generation of the ICFPM Model	124
8.2	Validation of the Prototype	125
8.2.1	Statistical Accuracy Testing	125
8.2.2	Scenario-based Testing	127
8.2.3	PPDM Scenario Test	135
8.2.4	Architecture Capacity Testing	136
8.3	Observations Gleaned regarding Fraudulent Claims	137
8.4	Discussion	138
9	Conclusion	139
9.1	Addressing the Problem Statement	139
9.1.1	The Main and Secondary Research Questions	139
9.2	Main Contributions	142
9.2.1	Advancing the State of the Art	142
9.2.2	Publications	143
9.3	Future Research	143
	References	145
	Appendix	173
	Appendix A: Example Code used to Predict Fraud	173
	Appendix B: SQL Code used to Generate Logistic Regression Test Data	175

Appendix C: Unsupervised Machine Learning Rules	177
Appendix D: Supervised Machine Learning Rules	177
Appendix E: OpenRefine API Call to Remove Empty Rows	177
Appendix F: OpenRefine API Call to Confirm DateOfClaim is a Date Value	178
Appendix G: OpenRefine API Call to Confirm AmountPaid is a Numeric Value	179
Appendix H: OpenRefine API Call to Confirm PolicyStartDate is Before Da- teOfClaim	180
Appendix I: OpenRefine API Call to Confirm ExcessPaid is Non-negative . .	181
Appendix J: Architecture Capacity Testing Results	182

List of Tables

2.1	Insurance Modernisation Methods	19
2.2	List of Fields Required to Describe a Claim	23
2.4	Issues in the Insurance Industry	27
3.1	Similarities and Differences between the Examples of Insurance Claims Fraud	34
3.2	The Occurrence of Fraud in the Claims-handling Process	40
6.1	Description of Variables Necessary to Predict Fraud Using this Re- search's Approach	92
8.1	Table Showing Data Used for Creating the F-test	126
8.2	Table Showing the Requirements of an F-test	127
8.3	Table Showing Field Values in Training Data-Set to Prove Scenario 1	129
8.4	Table Showing Subset of Rules for Scenario 1	130
8.5	Random Sample of Association Rules	131
8.6	Five Number Summary of Fabricated Claims Passed Through Logistic Regression	134
8.7	Five Number Summary of Actual Claims Passed Through Logistic Re- gression	134
1	Table Showing Results of Capacity Testing	182

List of Figures

1.1	Scope of the Study Reported on in this Dissertation	6
1.2	Layout of this Dissertation	11
3.1	Insurance Processes with Sub-processes derived from Saporito (2015) and Catlin et al. (2015)	36
3.2	The Insurance Claims Handling Process derived from van Jaarsveld et al. (2015) and Olalekan Yusuf and Ajemunigbohun (2015)	39
4.1	The Overlap of Data Science with Big Data	54
4.2	Big Data Science Component Diagram Derived from Dhar (2013), Rus-som et al. (2011) and Loukides (2011)	62
4.3	The Framework, Model and Architecture of this Research	72
5.1	The Necessary Components of a Software Framework from Research by Kopper (2009) and Phan et al. (2001)	76
5.2	Component Diagram of the Proposed Framework for Intelligent Insur-ance Claims Fraud Prediction	78
5.3	High-Level Use Case Diagram Showing the Claims Prediction Process .	79
6.1	Sequence Diagram Showing the Structure of this Chapter	84
6.2	Sequence Diagram Showing the Submission of a Claim and Application of Knowledge to this Claim	93
6.3	Chart Showing $W = W_0 + A(1 - e^{-kt})$ as derived from Marriott (2013)	96
6.4	Chart showing $a = 1 - e^{-0.597837t}$ as derived from Marriott (2013) . . .	97
6.5	Chart showing $b = 1 - e^{-0.162519t}$ as derived from Marriott (2013) . . .	98
6.6	Sequence Diagram Showing Generation of a Solution that can Intelli-gently Predict Insurance Claims Fraud	104
7.1	The Structure of the System to be Populated with Software and Hardware	111
7.2	Software Architecture of the Solution that Intelligently Predicts Insur-ance Claims Fraud	113
7.3	Hardware Architecture of the Solution that Intelligently Predicts Insur-ance Claims Fraud	114
8.1	Software Architecture of the Solution that Intelligently Predicts Insur-ance Claims Fraud	115

8.2	Adding Policyholder Information to the Prototype Insurance Claims Management System	117
8.3	Adding Policy Information to the Prototype Insurance Claims Management System	118
8.4	Adding Claims Information to the Prototype Insurance Claims Management System	118
8.5	Adding Third-Party Information to the Prototype Insurance Claims Management System	119
8.6	Web Page to set the Maintenance Variables	119
8.7	Dialogue Shown When Claim is Flagged as Potentially Fraudulent . . .	120
8.8	Insurance Claims Fraud Prediction Model	124
8.9	Maintenance Variables	128
8.10	Inputted Fields for the Claim with Fundamental Issues	131
8.11	Unsupervised Rules Broken for the Claim with Fundamental Issues . . .	132
8.12	Logistic Regression Results	133
8.13	Text Cleaning with the Levenshtein Algorithm	136
8.14	Architecture Capacity Testing: Number of Records vs Time Taken to Train the Model	137

1 Introduction

There has been an increase in the cost of insurance on a global scale and insurance fraud is identified as a contributing factor to this situation. Many cases of fraud are being reported, and the most common examples include arson by policyholders, faked car accidents, self-inflicted disability and over-inflated claims (CAIF, 2016; Viaene and Dedene, 2004). From this list, a trend can be noticed among cases of fraud, namely fabricating or incurring a loss (arson, faked accident or self-inflicted disability) to result in a claim and eventually in claim settlement. Although over-inflated claims do not imply that the loss was fabricated or incurred intentionally, parts of the claim are fabricated so that its monetary value can be increased.

Research into the causes of insurance fraud goes as far back as 1945, when Manes (1945) stated that the main cause of insurance fraud was financial profit. More recent research suggests that one of the causes of insurance fraud is a decrease in property prices, which has resulted in an increase in arson (Eriksen and Carson, 2017). Policyholders seem to believe that it is acceptable to make up their past premiums in current claims (Josephson, 2009) and owing to the large financial benefit of insurance, insurance fraud is becoming increasingly popular among organised crime syndicates (Laffey, 2004).

According to Ho (2014), the misappropriation of premiums is the most common type of insurance fraud. This implies that brokers or agents are commonly involved in insurance claims fraud.

The Horse Murders Scandal is one of the most infamous examples of insurance claims fraud that occurred in the USA in the property and casualty domain. Between the mid-1970s and 1990s, many show horses were insured for up to \$500,000.00 each. These horses were subsequently killed to collect insurance claims settlements and it is estimated that up to 100 horses were killed for this purpose (Nack, 1992). This example again shows that a claim was fabricated by deliberately incurring a loss to collect a claim settlement.

The cost of insurance fraud in America is estimated at \$80 billion per annum (Jordon, 2016), while in South Africa (a developing country) it is estimated at \$600 million (R8 billion) per annum. Insurance claims fraud makes up a considerable portion of these figures (Sera, 2016). It is estimated that 10% of all property and casualty insurance claims contain fraudulent elements (Insurance Information Institute, 2017). According to BusinessTech (2015), it is estimated that the average insurance company in South

Africa annually receives around 130,000 insurance claims. Owing to these large numbers, it may be difficult to give each claim the required attention in order to proactively predict fraud. These statistics also suggest that an intelligent method of predicting insurance claims fraud should add much value to the insurance industry. Therefore, it is essential to gain a clear understanding of intelligent financial fraud detection and prediction techniques. The difference between detection and prediction is that detection is used to uncover a fact, whereas prediction is the use of prior knowledge to state that something is going to happen (Oxford English Dictionary Online, 2017g,c).

According to West et al. (2014), intelligent financial fraud techniques have developed from traditional methods (such as the auditing of data in an attempt to uncover fraudulent transactions) to computational methods (based on statistics and artificial intelligence). The auditing of data to uncover fraudulent transactions can be seen as **detection**, as it is performed *after* the fact to uncover the fraudulent transactions. When an intelligent method is used to **predict** insurance claims fraud, the variables relating to a claim that has been submitted could be used to predict whether the claim is fraudulent or not (Liu and Chen, 2012). The advantages of using artificial intelligence include the fact that the latter can reduce operational time as it can mimic staff's decision making and ensure that systems remain relevant as time progresses (Chowdhury and Sadek, 2012). Examples of using artificial intelligence could be mimicking the rejection of claims (as would usually be done by a claim administrator), as well as adapting to claims adjusters behaviour. A disadvantage of artificial intelligence includes the fact that if a specific scenario has not been mimicked from the training data set, the artificial intelligence system will probably not react as it should.

A system may also act incorrectly if the parameters used to train the system have been chosen incorrectly (Chowdhury and Sadek, 2012) and, for example, rejected claims based on factors that have nothing to do with insurance fraud prediction. The advantages of using statistics-based computational methods include being able to show statistical measures of the fraud likelihood, as well as identifying new types of insurance fraud that had not previously been known (Li et al., 2008). Disadvantages of using statistical techniques are similar to the disadvantages of using artificial intelligence, in that statistical techniques can be difficult to train – which may result in the system not reacting the way it should and eventually lead to incorrect indications of insurance claims fraud (Li et al., 2008). Statistical and artificial intelligence techniques can be seen as part of the field of data science (Wang and Gu, 2016).

Data science should not be detached from domain knowledge (Waller and Fawcett, 2013). What this means, is that data science cannot replace domain knowledge and as such, it is more beneficial to use the two information sources in conjunction. Domain knowledge is a person's knowledge of a subject (Wildemuth, 2004) and in the case of claims handling, it is the knowledge that an experienced claims handler has acquired to identify suspicious claims. Therefore, it would be advantageous if claims handlers

combined their acquired knowledge of suspicious claims with a system that uses data science to process claims and to perform intelligent fraud prediction. Data science also has many other advantages. For instance, using sub-fields of data science (such as machine learning) results in better understanding of claims information and how the data in this information is interrelated, as one does not need to make assumptions about variables and their interrelationships (Dhar, 2013). West et al. (2014) maintain that manually auditing transactions can be time consuming, due to the onslaught of Big Data in the financial services industry. The fact that aged fraud identification techniques have become protracted and laborious is a valid point as the number of policyholders in South Africa was expected to be 16.2 million in 2017 (PWC and Metcalfe, 2014) and many insurers had more than a million policyholders (Santam, 2017; Hollard, 2017). These huge volumes give rise to large data sets, which means that Big Data solutions need to be taken into consideration. Using Big Data solutions has the advantages of better marketing, client segmentation, business insight, increased productivity, consumer benefits and sales possibilities (Sagiroglu and Sinanc, 2013; Chen et al., 2014) – all of which can be seen to overlap with the advantages of data science.

Client segmentation, increased productivity, consumer benefits and business insight are the key advantages considered in my research report. (Increased marketing and sales do not have a direct effect on the current research.) It would be valuable to gain insight into the data regarding the insurance business that can subsequently be applied to segment clients and classify them into fraudulent and non-fraudulent categories. If this division is done in an intelligent manner, it could increase the productivity of insurance companies when it comes to claims handling. Since insurance fraud costs the insurance industry remarkable amounts of money, a decrease in cost could greatly benefit the consumer (policyholder).

One of the largest issues of Big Data and data science technologies is the effect that it has on privacy (Koorn et al., 2015). Keeping large sets of data in one location makes it attractive for external attackers (Sagiroglu and Sinanc, 2013). In South Africa, this risk to the consumer is offset by legislation such as The Protection of Personal Information Act (PoPIA). The PoPI Act restricts the use and storage of data that relates to a person (Luck, 2014) and it is comparable to the United Kingdom's Data Protection Act as well as the European Union Data Protection Directives (EUDPD) (Botha et al., 2017). All this legislation shows that it is necessary to respect clients' privacy when creating a solution that uses Big Data. It may possibly be seen as redundant to be mindful of the storage of insurance data and *not* the use of this data for data science. However, value is added if data science is used in conjunction with privacy preservation methods. Swire (1997) maintains that members of the public are more likely to accept the use of their personal information if it is for relevant processes and if fair practices are used when processing such information. Thus, if insurance companies

are mindful about how personal information is processed and if they use fair practices during this processing, policyholders will be more accepting of the use of their personal information. In the insurance domain, personal information can include details such as age, gender, address and identification number (Bauer et al., 2006; Sahl Andersen et al., 2011). These details can be seen as sensitive private data and should be treated as such (Kaufman, 2009). It is important to be mindful of private data – not only to comply with legislative requirements, but also to protect consumers from threats such as identity theft and fraud (Antón et al., 2010).

1.1 Purpose Statement

Addressing fraud in financial data is not new, however, there is a gap within current fraud research in the financial sector. This gap arises by not adhering to restrictions when addressing fraud in property and casualty insurance. Namely, there is a lack of research that combines investigating the *prediction* instead of *detection* of insurance claims fraud within the property and casualty insurance domain with two boundaries; insurers have large sets of data and privacy needs to be considered when predicting fraud.

The aim of this research is to design and implement an intelligent way of predicting insurance claims fraud. This is done within the two aforementioned key boundaries: firstly – insurers have large sets of data; and secondly – privacy is important when it comes to personal information.

The purpose of the research study reported on in this dissertation can therefore be formulated as follows:

Intelligent methods can be used to predict insurance claims fraud. This can be conducted within the bounds and constraints of privacy legislation and the fact that insurers have large sets of data.

Confirming the above statement through the design and implementation of an intelligent solution is the objective of this research.

1.2 Problem Statement

The problem that has to be addressed can be described as follows:

There is an increase in the cost of insurance globally and it is compounded by the advent of fraud, of which insurance claims fraud makes up a substantial portion.

It is therefore the intent of the researcher to investigate the development of an intelligent way to predict whether insurance claims are fraudulent. The fact that insurers have large sets of data must also be considered in this investigation. One of the constraints of the proposed research is that it requires the creation of a successful solution – within the bounds of privacy legislation – to restrict the ways in which an individual’s insurance data can be stored and the purposes for which it can be used.

1.3 Research Questions

Based on the above, the following research questions need to be answered. The primary question to be answered is:

What elements should a solution have so as to be utilised to intelligently predict insurance claims fraud?

This question can be answered by posing three sub-questions on past fraud detection and prediction techniques.

- *Sub-question 1: What existing techniques are used to detect and predict insurance claims fraud?*

Once an understanding of past techniques has been gained, insight into two intelligent techniques can be gained through answering the following sub-question:

- *Sub-question 2: Can new developments in Big Data as well as in data science help to predict insurance claims fraud?*

The proposed research must determine whether new developments in Big Data and data science can be used as part of the intelligent solution for predicting fraud. Those aspects of Big Data that can be used to determine insurance claims fraud must be investigated, as well as the sub-fields and techniques in data science that could possibly add value to such fraud prediction. Once sub-question 2 has been successfully answered, the following issue needs to be investigated:

- *Sub-question 3: Can a solution that limits how insurance data is stored and what it can be used for, be generated within the restrictions imposed by privacy legislation?*

To facilitate the use of Big Data and data science within the bounds of privacy legislation, the solution needs to be created and designed in such a way that the privacy of neither policyholders, brokers nor insurers are compromised in this process.

Answering the above research questions and sub-questions should result in a better understanding of the intelligent prediction of fraud by using Big Data and data science technologies.

1.4 Scope of the Study

The scope of this study is depicted in Figure 1.1. The diagram starts by showing the possible scope of this research, but refines it by showing in green what falls inside and in red what falls outside the scope of the study in hand. The different types of insurance were derived from Baranoff et al. (2012).

To understand the diagram, it is important to note that property and casualty insurance is also known as short-term insurance or general insurance (Surbhi, 2017). The diagram shows that this study focuses on “Intelligently Predicting Insurance Claims Fraud”. Owing to the varying nature of insurance types, the scope of this study was limited to “Property and Casualty Insurance”. It was however not limited to a particular type of property and casualty insurance such as motor or household. The study includes different types of claims but was not further expanded to include other types of insurance such as “Life/Health Insurance”. Although the focus in my research was on a *developing* country, the findings may also be applied to the *developed* world.

The research was limited to “Intermediated Insurance”, which involved insurance brokers and agents, and it did not include “Direct Insurance”.

The diagram in Figure 1.1 is by far not exhaustive – it does not show all the possible types of insurance (“Other Types of Insurance”) as the insurance industry is vast.

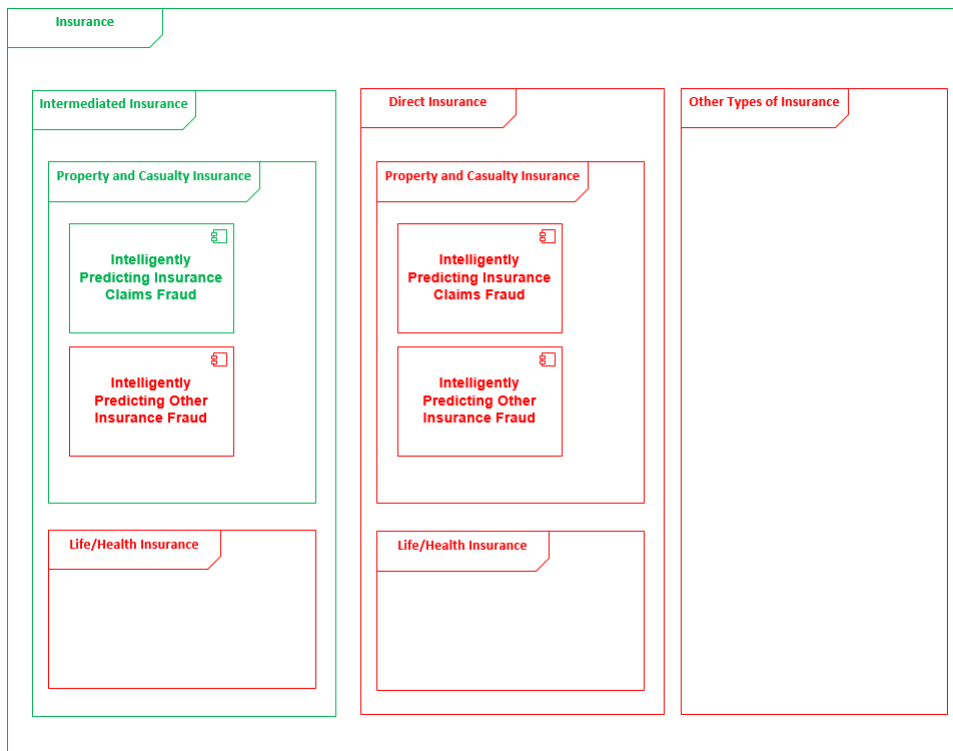


Figure 1.1: Scope of the Study Reported on in this Dissertation

1.5 Research Methodology Followed

In terms of the problem statement, the following four steps were taken to obtain the required research results.

1.5.1 Conducting a Literature Review

A literature review was conducted to facilitate a better understanding of the insurance industry and its acute data needs. The review was further expanded to gain an understanding of fraud in the insurance industry – with a focus on insurance claims fraud. A comprehensive list of existing techniques used in fraud prediction was obtained (including intelligent techniques), and a review was conducted of Big Data and data science technologies.

1.5.2 Developing a Research Use Case

Next, a research use case was developed that could be used to define the processes required to successfully test a possible solution. This research use case had to describe all tasks, systems and actors involved in predicting insurance claims fraud. Based on this list, the requirements were established of an intelligent solution that can be used to predict insurance claims fraud, as this is the focus of this research.

1.5.3 Designing and Developing a Solution

The third step that was followed in the research methodology was to design a solution based on research, and to develop it by taking into account the details, framework, architecture and a possible solution depicted by a model.

1.5.4 Validating the solution

Once the intelligent fraud prediction solution was ready, the final step was to validate its effectiveness by testing the prototype. The testing was done using test cases, examples and analysis.

1.6 Terminology

To understand the problem domain, it is necessary to have a better grasp of the terminology used in this dissertation. Hence, the researcher provides the definitions of the following key concepts and related terms involved in developing an intelligent solution to predict insurance claims fraud.

1.6.1 Cyber Security

Cyber security is the “collection of tools, policies, security concepts, security safeguards, guidelines, risk management approaches, actions, training, best practices, assurance and technologies that can be used to protect the cyber environment and organization and user’s assets” (Von Solms and Van Niekerk, 2013).

From this definition, cyber security, can be seen to prevent cyber crime.

Cyber crime are “offenses that can only be committed using a computer, computer networks or other form of information communications technology (ICT)” (McGuire and Dowling, 2013).

1.6.2 Intelligent Fraud Prediction

Since the aim of this research is to investigate an intelligent method of predicting insurance claims fraud, it is necessary to define “intelligent” and “intelligent fraud prediction”. These terms are evaluated in Chapter 3 of the research in hand.

Intelligent has many definitions but for the purpose of this research it is defined as:

Intelligent is “the ability for an information processing system to adapt to its environment with insufficient knowledge and resources” (Wang, 2007).

This definition fits the purpose of this research, as machine learning and statistical methods are shown in Chapter 4 to be intelligent by having the ability to adapt to an environment and learn from data provided. It is shown in Chapter 5 that Apriori association rules and logistic regression will be used to predict insurance claims fraud and hence this is intelligent with regards to this research.

A definition for intelligent fraud prediction could not be found in credible dictionaries such as the Oxford English Dictionary Online (2017k) and Cambridge English Dictionary (2017a), as they did not yet contain such definitions at the time of this research. The researcher therefore derived a definition from the definitions of “intelligent”, “fraud” and “prediction” in Chapter 3 of this research. For the purpose of this research, intelligent fraud prediction is defined as follows:

Intelligent fraud prediction involves using a system that can adapt to the environment using understanding but lack of knowledge to pre-emptively determine whether a person has achieved an unjust advantage over another.

1.6.3 Insurance Claims Fraud

Insurance fraud is “the wrongful or criminal deception of an insurance company for the purpose of wrongfully receiving compensation or benefits” (Legal Dictionary, 2016).

Insurance fraud can be further split into:

- **Hard (planned) fraud:** Criminals who fabricate transactions, accidents or injuries are seen as committing planned (hard) fraud.
- **Soft (opportunistic) fraud:** Policyholders who mislead insurers with an over-inflated claim amount to increase their financial gain can be seen as committing opportunistic (soft) fraud (Goel, 2014; Tennyson, 2002).

Because there are multiple variations of fraud in insurance (Wells, 2013), insurance claims fraud must be further specified as follows:

Insurance claims fraud is the occurrence that an individual(s) exaggerates or falsifies claims information to receive compensation from an insurance company (USAA, 2017).

1.6.4 Big Data Science

Big Data can be described as “high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation” (Gartner, 2016).

Data science is the technological management, acquisition and analysis of data (Hardin et al., 2015).

Predictive analytics is “the use of statistical or machine learning methods to make predictions about future or unknown outcomes” (Brown et al., 2015).

Having large data sets does not add much value, unless enhanced insight can be gleaned from it.

The convergence or overlap among these three fields, namely Big Data, data science and predictive analytics (Waller and Fawcett, 2013), is for the purposes of this research referred to as Big Data Science (BDS). This term will be expanded upon further in the dissertation.

1.6.5 Privacy

Because this research has a focus on privacy, terms such as information privacy, personal information, anonymous data and pseudonymous data need to be defined to better understand the research.

Privacy with regard to the current research involves the control of information flow during any phase of information processing, use, acquisition and disclosure (Finn et al., 2013).

Personal information is seen as any information that can be identified to relate or refer to an individual (Al-Fedaghi, 2006).

Anonymous data is data in which transactions in the data set cannot identify a person – either by a singular variable or by a combination of variables (Clarke, 1999).

Pseudonymous data is data in which transactions in the data set cannot identify a person through normal methods, but only when specific additional data is associated with it (Clarke, 1999).

1.6.6 Solution

The solution designed and developed in this dissertation included a framework, architecture and model. These components of the solution need to be defined, with regards to this research, to prevent misunderstanding.

An **Architecture** is the “conceptual structure and overall logical organisation of a computer or computer-based system from the point of view of its use or design; a particular realisation of this” (Oxford English Dictionary Online, 2017b).

A **Framework** defines the interfaces, rules and services that are required for it to perform a set of tasks (Kopper, 2009).

A **Model** is “A simplified or idealized description or conception of a particular system, situation, or process, often in mathematical terms, that is put forward as a basis for theoretical or empirical understanding, or for calculations, predictions, etc.” (Oxford English Dictionary Online, 2018).

1.7 Layout of this Dissertation

Figure 1.2 shows that this dissertation contains ten chapters. They are as follows:

Chapter 1 supplies an introduction to the dissertation as well as a description of its structure. It explains the terminology commonly used in the dissertation and provides some background about the purpose of the research.

Chapter 2 provides an overview of the insurance industry. This overview focuses on data, the data needs that have arisen and the technologies that have been developed to meet these needs.

Chapter 3 examines the occurrence of fraud in the insurance industry. A sub-set of fraud, namely insurance claims fraud, is further examined. The chapter also examines the current and past research performed with regard to fraud prediction in financial systems.

Chapter 4 gives an overview of Big Data Science and describes those aspects of data that changes it into Big Data. It also describes the field of data science and how this is coupled with Big Data.

Chapter 5 describes a framework for the solution proposed in this research and discusses a high-level use case of this solution aimed at intelligently predicting insurance claims fraud during the claims-handling process.

Chapter 6 exposes the detail of the high-level design by suggesting a complete process of predicting insurance claims fraud. The chapter also includes a proposed model that can be used to intelligently predict insurance claims fraud.

Chapter 7 puts forth a high-level architecture that can be used to facilitate the process when the goal is to intelligently predict insurance claims fraud.

Chapter 8 constructs and validates the prototype by exposing scenario-based results as well as the results of a statistical test.

Chapter 9 concludes the dissertation.

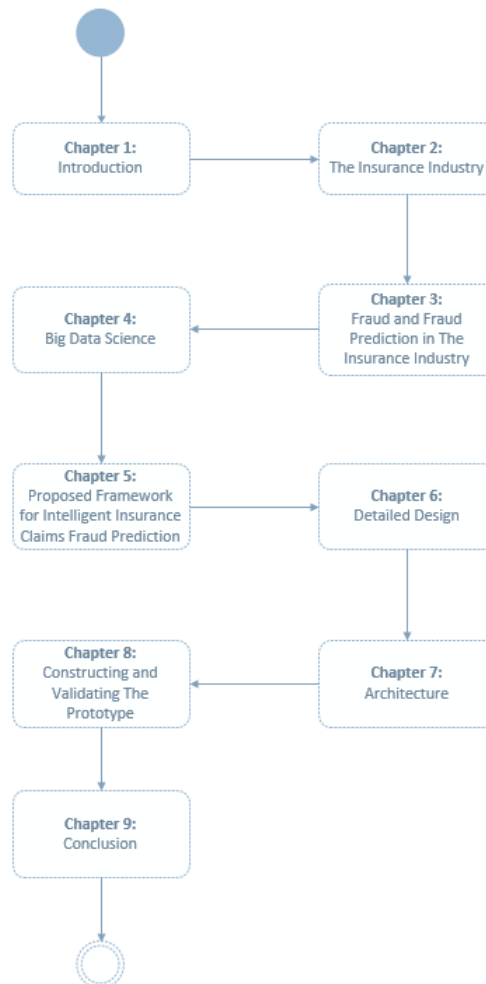


Figure 1.2: Layout of this Dissertation

2 The Insurance Industry

The purpose of this research as described in Chapter 1 was to determine an intelligent way of predicting insurance claims fraud. Thus it is important for the reader to gain a clear understanding of the insurance industry.

No single definition of ‘insurance’ exists as yet, and to date there are many conflicting opinions in the field (Talesh, 2015). The common theme that runs through most of these opinions is that insurance is a structured, regulated method of splitting costs for detrimental experiences. Insurance is therefore based on the chance or risk that a person, enterprise or tangible resource will have injurious harm come upon them. The purpose of insurance is to reduce the effect of the damage after the harm has occurred (Abraham, 2012). Modern insurance practices are arranged into a formal process for dispersing cost among people and organisations if a risk comes to fruition. Ideally, the loss is provided for in advance, by charging a premium (Talesh, 2015).

Chapter 2 describes the current state of the insurance industry from the researcher’s perspective; this is done to provide a better understanding of what is required due to the issues in the industry. In particular, the research focuses on existing perspectives of data such as data mining, Big Data and data science. The approach adopted in this chapter focuses on these three data areas, as insurance covers a vast field. The chapter describes data in insurance and suggests a data standard that is key to the generic application of this research globally. Data standards provide a standard way of representing a data set (Jensen and Shumway, 2010). This is important for this research to ensure that its findings can be applied to any insurance company that adopts the standard.

2.1 Issues in the Insurance Industry with regard to Data

Of the many issues currently facing the insurance industry, the acute data needs of the insurance industry are discussed in this section, with a specific focus on data science and advanced data analytics.

A study conducted by PriceWaterhouseCoopers (PWC) in 2015 to determine the Top Issues in the Insurance Industry (Trowbridge and Rose, 2015) culminated in a report

that is comparable to the reports by Mercer (2014) and Ernst & Young (2013). Of the three reports, Trowbridge and Rose's (2015) was the most current and it related most to data issues in the insurance industry. A more recent version of the PWC report (Trowbridge and Rose, 2017), which deals with similar issues in the insurance industry (e.g. the impact of technology, taxation, operating models and insurance deals) was not available at the start of the research, but the themes in both reports are similar. In the 2015 PWC report, which was used to structure sections 2.2.1 to 2.2.4, PriceWaterhouseCoopers named strategies, taxation, operations, regulations and insurance modernisation as the five top issues in the insurance industry that generated acute needs (Trowbridge and Rose, 2015). In the current research, four of these issues – strategies, operations, regulations and insurance modernisation – were found to have a direct impact on insurance data needs. These issues are analysed below with alternate literature contributions:

2.1.1 Strategies

Halverson and Malhotra (2015) argue that insurance companies need to focus their strategy on partnering with technology companies to develop new services as well as insurance products.

An example of a strategy that should be adopted is assessment of the potential impact of automated driver assistance systems (Bengler et al., 2014; Gschwendtner et al., 2014). According to Moore and Zuby (2013), certain collision avoidance systems can result in claim reductions due to fewer accidents. The impact of automated driver assistance systems that can use Artificial Intelligence (AI) techniques needs to be factored in when determining risk and predicting claims.

Another key strategy adopted in the insurance industry is group insurance. One of the main reasons why group insurance has increased so much in stature was the development of Big Data (Trowbridge and Rose, 2015). There is now an abundance of information that was not previously stored by insurance companies because manual processes were in place to capture the information. The mammoth growth in data has given rise to analytics that could possibly break down the silos that have formed in the insurance industry (Trowbridge and Rose, 2015).

2.1.2 Operations

The main change introduced in operations involves upgrading legacy insurance systems and processes (Ernst & Young, 2013; Trowbridge and Rose, 2015). The new systems include features such as Big Data analytics, improved efficiency, advanced underwriting and optimised policy pricing (Trowbridge and Rose, 2015). Operations needs to be

upgraded to work with these systems and improve their billing, underwriting, policy administration and claims handling (Ernst & Young, 2013).

2.1.3 Regulations

According to Trowbridge and Rose (2015) the CEOs of insurance companies are nowadays spending far more time and effort on the regulatory requirements of the insurance industry and they are adopting new methods to meet regulatory needs. One method being adopted by business to meet regulatory needs is data-driven decision making, which uses data and information to drive the questions and answers regarding regulatory changes. Data-driven decision making is also important for the regulatory bodies that draft legislation (Schneiberg and Bartley, 2001).

An example of a regulation that will affect the strategies of insurance companies is the International Capital Standard (Trowbridge and Rose, 2015). This regulation involves creating a standard method of reporting the capital of an insurance company. It also requires insurance companies to change their systems and models and once again, it is noticeable that the basis of these systems would be data (Essert and Barron, 2015).

From the definition of regulations, namely “principles and rules that govern practice as well as behaviour” (Oxford English Dictionary Online, 2017i), it is evident that policyholders can be regulated as well. Hence insurance fraud can be expected to be affected by regulations. Van Wolferen et al. (2013) maintain that there is currently uncertainty about the extent of fraud in the insurance industry as the available statistics are not accurate. If insurance companies were required to publish the results of audits implying that fraud had been committed, it would perhaps be easier to understand fraudulent trends. This is a regulatory requirement that could be solved by data.

The needs of the insurance industry in developing countries such as South Africa are similar to those in the global insurance industry. Regulatory requirements that are specific to South Africa are comparable with similar legislation in both developing and developed countries. These requirements have facets that are directly related to data. To gain an understanding of such regulatory requirements, The Retail Distribution Review (RDR), Solvency Assessment and Management (SAM) and PoPIA are discussed below. These regulatory requirements were mentioned by both Makhafola (2015) and Knoesen (2017) as issues in the South African insurance industry.

2.1.3.1 Retail Distribution Review (RDR)

The Financial Sector Conduct Authority (FSCA) (www.fsca.co.za), which is responsible for the regulation of non-banking financial services in South Africa, released a report intent on regulating financial services known as the Retail Distribution Review

(RDR) (FSB, 2014). The main purpose of the RDR was to compel financial service providers to adhere to the Treat-Customers-Fairly (TCF) Principles (Tomlinson, 2015). Many other objectives were outlined in the RDR, but the one that has probably had the greatest effect on Information Technology providers and departments in the insurance industry is the introduction of legislation that will force Insurance Intermediaries to keep customer data and to give advisers and product suppliers continued access to this data (Financial Services Board, 2015).

The RDR was meant to ensure that customers have access to advice that is of a high quality and not subject to a conflict of interests. To ensure a competitive advice that benefits the customer and to ensure that the advice supplied by advisers is compliant with RDR, could however prove to be an expensive task. To reduce this expense, Tomlinson (2015) suggests the use of systems that can provide the advice with little input from costly advisers. For example, if systems can provide underwriting pricing by using large sets of data (Geraghty, 2001), a financial adviser would not be required to do the work. This notion agrees directly with the research of Andrews (2013), which shows that Big Data and Advanced Analytics add great value to the insurance industry.

In the current research, these technologies were examined in many different examples and as such, were incorporated into a solution that can be used to predict insurance claims fraud. High-quality advice would need to be given to policyholders and because the source of such advice would be data, more data would have to be kept by insurers. This would add value to the prediction of insurance claims fraud, as a far greater source of data would now be used to generate an intelligent solution to counter such fraud.

2.1.3.2 Solvency Assessment and Management (SAM)

According to Martin and Hayes (2013), Solvency Assessment and Management (SAM) is intended to address the operational risk problems currently facing the insurance industry. SAM is intended to align insurers in South Africa with the Solvency II directive of the European Union (Eling et al., 2007). This requires insurers to prove that they can remain solvent throughout all possible quantifiable losses during a year. To do this, they will have to use either the Standard Formula Approach (EIOPA, 2014) or an Internal Model Approach. The Standard Formula Approach offers simple lines whereas the Internal Model Approach is much more complex. All internal models will need to be approved by the Financial Sector Conduct Authority. Internal models can use a Loss-Distribution Approach (LDA) (Frachot et al., 2001) which is driven using data. Frachot and Roncalli (2002) maintain that when both internal and external data sets are used to work out operational risk, the results are much more accurate than when only internal data was used. Once again, it is noticeable that different sources of data are being used to solve issues in the insurance industry. Although this does not correlate directly with fraud prediction, the use of data to solve an issue is self-evident.

If insurers have to store more data as a SAM requirement, this data can also be used to predict insurance claims fraud in an intelligent manner.

2.1.3.3 PoPIA

Due to the fact that Big Data and advanced analytics involve the sourcing of data from previously inaccessible places, new social and ethical issues have arisen. Some researchers (Swedloff, 2014; Crawford, 2011) warn that the use of Big Data in the insurance industry can be unfair or unethical and may at times unknowingly reinforce stereotypes that were previously avoided. Big Data usage has an effect on privacy, as data that is used for policy pricing may possibly have been sourced from repositories without the customer knowing (Swedloff, 2014).

The South African Government nonetheless introduced legislation that aims to protect people against an invasion of their privacy, namely The Protection of Personal Information Act (PoPIA) (Luck, 2014). PoPIA regulates how South African citizens' data should be stored and processed. People have to be notified when their data is going to be stored, and their consent needs to be obtained. If insurance companies (for example) do not adhere to these regulations, they could lose customers, damage their reputation and face fines or jail time (Botha et al., 2015). PoPIA prescribes eight conditions in respect of data privacy, as listed below (De Bruyn, 2014; Red Edge Solutions, 2015):

- The collector of the personal information is held accountable for adherence to the PoPI Act.
- The collection of the personal information must be fair and lawful.
- During the collection of personal information, the subject needs to be notified and informed about the purpose(s) for which the data is being collected.
- Personal information may be used only for what was specified during notification.
- Personal information needs to be accurate and have data integrity.
- The subject of the personal information needs to be aware of the collection of his/her personal information.
- The collector of the personal information needs to impose data security safeguards.
- The subject of the personal information has the right to view his/her personal information and change or delete it.

These conditions could possibly offset the effect of the use of Big Data and data science as the process needs to be more transparent. The condition that requires notifying the subject of the use of his/her personal data and what it will be used for cannot be offset through this research. However, if an insurer actually informs policyholders on taking out a policy that their data will be used in an intelligent manner to predict insurance

claims fraud, then this research can be applied. The research can however be affected by the fact that the data needs to have data security safeguards. The personal information would need to be protected in an intelligent solution that is used to predict fraud. This legislation can be compared to the European Union's Data Protection Directive and the United Kingdom's Data Protection Act (Botha et al., 2015). As PoPIA is newer than other such legislation, concepts such as anonymous and pseudonymous data have not arisen (European Union, 2016). According to Floridi (2014), data is considered anonymous if the person to whom the data relates cannot be re-identified from the data. If data is anonymous, then the data could possibly be kept past the point where a policyholder asks for his/her personal information to be removed. Since these terms are not contained in the legislation, they will need to be explored with the legislative bodies when the legislation comes to fruition. Such privacy legislation has a direct effect on this research and, as such, the focus of this research is on protecting privacy while intelligently predicting insurance claims fraud.

2.1.4 Insurance Modernisation

According to Galaez et al. (2015), data in the insurance industry has many issues. For instance, data originates from legacy systems, engines and platforms. Furthermore, Information Technology departments do not understand the business needs to perform analyses. The data is also not structured and it is poorly managed. Since there are so many sources, this results in conflicting information and it is costly to ensure that data needs are met. The issues mentioned indicate that there is a need for insurance modernisation. Systems should be upgraded to meet the needs of insurance companies.

Besides these issues, one of the main strategies of insurance modernisation is the use of advanced analytics (Bosco, 2014). According to Bose (2009), advanced analytics is a group of techniques used to solve issues and gain insight by analysing data to predict outcomes. This is a key component of insurance modernisation. It is also noticeable that there are new metrics that have come to the fore in recent years. The key focus of these new metrics are financial reporting, regulatory reporting and reporting that creates business value (Trowbridge and Rose, 2015). Redesigning the data and processes in insurance companies could result in lower costs. Of the many different applications of advanced analytics, some can be used to solve the needs of the insurance industry. The following gives an overview of possible examples of advanced analytics use within the insurance industry.

2.1.4.1 Advanced Analytics-Based Insurance Modernisation

Many types of advanced analytics are currently utilised in the insurance industry. From a comparison of research that describes examples of advanced analytics in insurance,

Boobier (2016), Maas et al. (2014) and Bharal and Halfon (2013) all mention the use of advanced analytics to improve underwriting, gain customer insight and detect fraud. It is for this reason that these three insurance modernisation methods that use advanced analytics are described in Table 2.1. A key question (see sub-question 2 in Section 1.4) to be answered in this research is “*Can new developments in Big Data as well as in data science help to predict insurance claims fraud?*”. Seeing that there is an overlap between advanced analytics, predictive analytics, data science and Big Data (Waller and Fawcett, 2013), the insurance modernisation methods described in Table 2.1 are important to understand (Big Data and data science are currently used in the insurance industry).

Table 2.1: Insurance Modernisation Methods

Insurance Modernisation Method	Description
Gaining customer insight through advanced analytics	According to Maas et al. (2014), using Big Data, data science and advanced analytics can help insurance companies to gain insight into their customers and provide new products and services based on this insight. Gaining customer insight can also result in the retention of customers (Maas et al., 2014). Using data science and analytics assist insurance companies to understand better how to retain their customers (IBM, 2010).
Advanced analytics to solve complex underwriting pricing	To stay profitable and relevant in years to come, using Big Data in underwriting is key for insurance companies (Andrews, 2013). Insurers believe that they do not have the expertise or facilities to leverage Big Data in their underwriting. The underwriters of insurance premiums consider this a shortfall, as having access to real-time claims data and insight gained by means of by Big Data analytics is key to determining risk accurately (Andrews, 2013). Siegelman (2014) maintains that the use of Big Data and data science-based analytics allows insurance companies to utilise a wider range of variables and factors when determining risk – which can result in the increased accuracy of underwriting.
Advanced analytics for the prevention of insurance fraud	<p>Big Data, data science and advanced analytics can also be used to detect fraud in the claims management process (Maas et al., 2014). Data science methods such as data mining can be used to determine patterns and extract information based on these patterns. Techniques such as Bayes technique, decision tree algorithms, support vector machines, clustering, prediction, outlier detection, regression and visualisation can be used in the data mining process to predict insurance claims fraud (Sithic and Balasubramanian, 2013).</p> <p>There seems to be conflicting information as to the benefit of using data science in the prediction of insurance fraud. Ormerod et al. (2003) maintain that using techniques such as data mining can produce a high chance of false positives. Innocent customers can be pursued, while guilty fraudsters are missed due to the fact that they know which indicators would show if they committed fraud. Another problem is that highly skilled people are still needed to interpret the results of analysed insurance data to predict fraud (Ormerod et al., 2003). Although this seems to minimise the benefit of using advanced analytics to detect fraud, it has to be admitted that finding patterns and inconsistencies in insurance data can be incredibly beneficial to the industry.</p> <p>Using Big Data and data science in the prevention of insurance fraud is an intelligent way of predicting insurance claims fraud and should be investigated further. Therefore, this outline cannot suffice and will be expanded upon throughout this research.</p>

Common patterns can be derived from the three insurance modernisation methods that use advanced analytics, and relating these patterns to the purpose of this dissertation could add value. One of the research sub-questions posed in Section 1.4 was: “*Can new developments in Big Data as well as in data science help to predict insurance claims fraud?*”. Of the three examples of insurance modernisation methods that use advanced analytics, each example mentioned the use of Big Data as well as data science as part of the modernisation method. This comparison reveals that Big Data and data science are commonly used when performing advanced analytics in insurance. Big Data does not necessarily need to be used to perform advanced analytics, but some method of data science needs to be used. Having access to Big Data does not add much value unless enhanced insight can be gleaned from it.

Another research question posed as Sub-question 3 was formulated as follows: “*Can a solution that limits how insurance data is stored and what it can be used for, be generated within the restrictions imposed by privacy legislation?*”. This question was necessary because each of the three examples described in Table 2.1 use multiple sources of data in their advanced analytics and these sources of data would include personal information.

Big Data, data science and personal information are key resources in insurance modernisation.

2.2 Data in the Insurance Industry

Data is a driving force to solve the issues associated with problems in the insurance industry. It is therefore valuable to gain an understanding of the data in the insurance industry.

The basis of insurance processes today is data. The insurance industry currently uses two types of data – internal data or external data (Sadiq et al., 2004). Internal data includes risk data and accounting data. Risk data can be facts such as premium data, policy data, exposures, losses, claim counts and explanatory facts about policies or claims, whereas the accounting data include underwriting expenses and claim expenses (Werner and Modlin, 2010). The external sources of data on the other hand include statistical plans, aggregated insurance industry data and other third-party data (Werner and Modlin, 2010). Statistical plans are either summary-based or transaction-based aggregated historical insurance data that is compiled by regulators (Kuys and Zehnirith, 1997). Aggregated insurance data is similar, but it is compiled by various organisations and not for regulatory purposes. An example of such an organisation is the Insurance Institute for Highway Safety in the US, which collates data from various insurers to

provide ratings for motor vehicles (IIHS, 2016). Something similar is performed by The South African Insurance Crime Bureau, which collates fraudulent activity data in South Africa (SAICB, 2012).

In contrast to internal data, third-party data is not insurance-specific and can include geo-demographic data, public rate filings, commercial credit scores, commercial auto segmentation and telematics or usage-based insurance data (Kolde and Walker, 2015).

Staples (2011) states that the data that the insurance companies do not leverage, but have access to, include customer data, consumer life-cycle data and marketing data. These can be seen as both internal and external sources of data. Companies do not leverage this data as they do not possess the expertise or the tools to analyse it. The type and quantity of data that is currently available to but not adequately used by insurers, brokers and intermediaries, provide opportunities for the insurance industry. These opportunities arise from the procuring of new customers, to retaining them and providing new services and products. In the case of the current research, this data could perhaps be utilised during the claims process to predict insurance claims fraud.

The internal and external data sources required during the insurance process are specific to the type of analyses being done (Werner and Modlin, 2010).

2.2.1 Data in the Insurance Industry regarding Data Quality Tools

Since data in the insurance industry originates from legacy systems, engines and platforms (Galaez et al., 2015), it is difficult to use the data. A solution to this problem can be to use any of the wide range of data quality tools, such as Informatica, RedPoint, Experian, MIOsoft, IBM, Information Builders, Wrangler or OpenRefine (Friedman and Judah, 2015; Krishnan et al., 2016). Examples of what these tools can do include (but are not limited to) data integration, matching, parsing, standardisation, profiling, cleansing, migration and data analytics (Informatica, 2015). According to Experian (2017), using data quality tools in the insurance industry can increase operational efficiency, improve service, create better risk assessment and prevent the loss of premium. As previously mentioned, privacy legislation such as PoPIA requires data to be correct and hence data quality tools could add value in this process. Data quality tools can also be used as part of an intelligent method of predicting insurance claims fraud to ensure the completeness and integrity of the data used. (This issue will be dealt with later on in the research.)

2.2.2 Data in the Insurance Industry regarding Data Standards

Data standards are vital to ensure that there is uniformity among data sets (Jensen and Shumway, 2010). This is important in this research as it will allow for the application

of its findings in any insurance company that uses the standards. According to Nath (2016), the fact that there is a communication disconnect among insurers results in criminals exploiting the lack of visibility. If insurers share data, this visibility will be increased and could provide a key benefit in the form of the reduction of fraud. To share it, however, the data would need to adhere to standards so that it could be understood by more than one insurer. The ACORD (Association for Cooperative Operations Research and Development) framework contains one such standard that is widely used and is a global standard, which means that the data is understood internationally (Lloyds, 2017). Since data needs to be shared by intermediaries and insurers in South Africa (FSB, 2014), a data exchange was set up by Astute, allowing insurers and intermediaries to share data if there is a common interest (i.e. if the policyholder is a client of both the insurer and the broker) (Astute, 2013). Astute (2013) bases the messages of the exchange on the ACORD framework.

ACORD, an organisation started by global insurance companies, aims to define the Enterprise Architecture of insurance companies. The ACORD framework, which is used as a basis for modelling insurance companies, their processes, systems and data (Krstajiić et al., 2014), includes the business glossary of terms, the capability model, component model and the data model (included in the information model) (ACORD, 2015). The ACORD framework also creates a standard way of specifying data and how to transfer this data within the insurance industry (ACORD, 2015). Mazur (2011) maintains that the ACORD framework is an integral part of the insurance industry as it is an XML standard that allows the integration of internal and external data. The framework specifies messages such as insurance applications as well as the transmission of underwriting requirements. The standards of the ACORD framework are flexible in their capability and application. This is both a positive and negative characteristic of the standards as they cater for most messages in insurance, but these messages can be ambiguous. For the purposes of this research, the “Data Model” would contain a data standard that could be used to create uniformity in the insurance industry and cater for the sharing of data between insurers and brokers. Although the sharing of data could reduce the onset of fraud, insurers are not likely to follow this route, as it could result in other insurers using their data for lead generation. The methods proposed for data sharing would have to take this into consideration and, once again, the protection of personal information needs to be considered.

2.2.3 Data in the Insurance Industry regarding Insurance Claims

Because this research focuses on insurance claims fraud, an indication as to the type of data contained in a claim could add value. For the current research, the possible fields in a claim were identified by collating what information should be kept by different sources. By viewing the ACORD standard (ACORD, 2015) many possible fields

became evident. These fields were filtered and adjusted so as to meet the requirements of the developing world with a use case in South Africa, and to meet the scope of the research. The scope has been applied to generic insurance claims, and both policyholder and claims-specific information were included. Information contained in the standard, such as “Watercraft” and “Vehicle” (ACORD, 2015), were excluded as this could limit the scope of the solution. Interviews with an insurance system provider (“Broker Management System Support Manager”, 2016) as well as an A-rated insurer (“General Manager: Risk and Technology”, 2016) were conducted. The discussion below outlines a non-exhaustive list of possible fields in a claim required for the research.

These fields include policyholder information such as personal and claims-specific information. It should be noted that if the fields required had to meet a certain standard (such as the ACORD Standard), they had to be labelled as such. The fields shown in Table 2.2 do not adhere to the ACORD naming standards in order to prevent an infringement of the intellectual property rights of ACORD. The field names are generic so that they can be mapped to ACORD fields. Table 2.2 shows the name of each suggested claim field with a corresponding description of the field. The table also shows a “Category” that determines the source type of the information, namely “Insurance Provider”, “Claim” “Policyholder”. Lastly, owing to the focus of this research on privacy, the table indicates whether the field can be regarded as personal information.

Table 2.2: List of Fields Required to Describe a Claim

Field	Description	Category	Personal Information
Agent	The agent employed by the brokerage to manage the policyholder’s policy	Insurance provider	X
Fraudulent claim reason	Reason why claim was considered fraudulent (not mandatory)	Claim	
Date of loss	Date that incident occurred, resulting in loss for policyholder	Claim	
Time of loss	Time that incident occurred, resulting in loss for policyholder	Claim	
Date of claim	Date on which policyholder reported loss to broker/insurer	Claim	
Agency/Broker’s unique ID	Field to uniquely identify the broker	Insurance provider	
Insurer’s unique ID	Field to uniquely identify the insurer	Insurance provider	

Policyholder's name	The name of the person holding the policy	Policy	X
Policyholder's surname	The surname of the person holding the policy	Policy	X
Policyholder's telephone no	The telephone number of the policyholder	Policy	X
Age	Age of policyholder	Policy	X
Gender	Gender of policyholder	Policy	X
Kind of loss	Type of loss that occurred (fire, theft, etc.)	Claim	
Incident address	Address at which incident occurred	Claim	
Address of policyholder	Residential address of policyholder	Policy	X
Police or fire dept. to which incident was reported	Name/ Area of Police/ Fire Dept. where injury or harm was reported	Claim	
Policyholder's street	Street address of policyholder	Policy	X
Policyholder's province	Province where policyholder lives	Policy	X
Policyholder's city	City where policyholder lives	Policy	X
Policyholder's area	Area where policyholder lives	Policy	X
Policyholder's postal code	Postal code of area where policyholder lives	Policy	X
Province	Province where loss occurred	Claim	
City	City where loss occurred	Claim	
Area	Area where loss occurred	Claim	
Postal code	Postal code where loss occurred	Claim	
Marital status	Marital status of policyholder	Policy	X
Date of birth	Date of birth of policyholder	Policy	X
Sum insured	The maximum amount of money that the insurer will pay to the policyholder	Policy	
Probable amount of entire loss	The policyholder's estimation of the value / expense of the loss	Claim	

Date of settlement	The date on which the claim was settled by insurer/ broker	Claim	
Total policies revenue	The total premium that the insurer receives from the policyholder	Policy	
Amount claimed	After evaluation, the value of the amount claimed	Claim	
Amount paid	The claim amount that the insurer pays to the policyholder	Claim	
Payment account number	The account number into which the claim was paid	Claim	X
Policy start date	The start date of the policyholder's insurance coverage	Policy	
Policy end date	The end date of the policyholder's insurance coverage	Policy	
Other party's damage	The amount claimed from a third party for incident	Claim	
Other party's name	Name of the other person if more than one person was involved in the incident	Claim	X
Other party's surname	Surname of the other person if more than one person was involved in the incident	Claim	X
Other party's insurer	The name of the insurer of the other party	Claim	
Other party's street	Street address of other party	Claim	X
Other party's province	Province where other party lives	Claim	X
Other party's city	City where other party lives	Claim	X
Other party's area	Area where other party lives	Claim	X
Other party's postal code	Postal code where other party lives	Claim	X
Assessor's unique identifier	Unique identifier of the claims assessor	Claim	
Total excess	The total excess to be paid by the policyholder	Policy	
Type of insurance coverage	The type of insurance coverage, e.g. comprehensive only	Policy	

Title holder's name	The name of the person who owns the property/items damaged/lost	Policy	X
Title holder's surname	The surname of the person who owns the property/items damaged/lost	Policy	X
Short description	Short description of the loss	Claim	
Long description	Long description of the loss	Claim	
Claim service provider	Service provider who will repair/offset the damage	Claim	

2.3 The Use of Data to Solve the Issues in the Insurance Industry

Although data is central to the insurance industry these days, many issues relating to data still need to be resolved. Whether it involves internal or external data, the power that such resolution can give to the insurance industry is immense. The issues in the insurance industry may not be data centric, but possible solutions can be found through the use of data. The sections below give an overview of the issues dealt with in this chapter. Recommendations to address them are also included.

2.3.1 Strategies and Data

The strategies mentioned in this chapter focused on insurance companies partnering with technology firms to develop new services and products. Examples of such services and products included utilising artificial intelligence (AI) techniques to determine risk when automated driver assistance systems were used, as well as strategically using Big Data to break down the silos in the insurance industry.

2.3.2 Operations and Data

The operational issues mentioned in this chapter involved the fact that insurance systems regularly included legacy systems that had to be upgraded. Upgrading these systems to include techniques such as Big Data Analytics could result in improved efficiency, advanced underwriting and policy pricing.

The operational departments of insurance companies would need to adapt and utilise these systems to improve service offerings.

2.3.3 Regulations and Data

Current and proposed legislation included the International Capital Standard, RDR, SAM, PoPIA and regulatory requirements regarding fraud. The study found that data-driven decision making was being utilised to drive regulatory changes. It was also determined that the regulatory changes would result in an increased amount of data being kept.

2.3.4 Insurance Modernisation and Data

The focus of insurance modernisation in this chapter was on the upgrading of legacy systems and the use of advanced analytics in insurance. Three examples of advanced analytics included gaining customer insight, the latter's application in complex underwriting pricing and its use for the prevention of insurance fraud. This was shown to be possible through the use of Big Data and data science.

2.3.5 Comparison of Issues with Possible Solutions

Table 2.4 lists all the issues and suggests possible solutions to remedy them. The issues were derived from the groups of solutions mentioned in this chapter. The solutions were generic and they did not contain details that were too specific to the problem, as more value would be added where there was overlap. The possible solutions included data cleaning, data science, Big Data, new sources of data and privacy considerations. To understand this table, one of the "Issues" can be described as follows: "Strategies" can be seen to provide a "Possible Solution" if they use "Data Science" in conjunction with "Big Data" to stay profitable and relevant in the insurance industry by developing new products and services.

Table 2.4: Issues in the Insurance Industry

Issues	Possible Solution				
	Data Cleaning	Data Science	Big Data	Privacy Considerations	New Sources of Data
Strategies		X	X		
Operations	X		X		
Regulations	X	X	X	X	X
Insurance Modernisation		X	X		X

Furthermore, Big Data and data science are key solutions to issues that the insurance

industry is currently experiencing. Data cleaning and privacy considerations can also be seen to add value and as such, need to be investigated with regard to finding an intelligent way to predict insurance claims fraud.

2.4 Discussion

Many data-centric issues are facing the insurance industry. These can possibly be resolved through solutions such as data mining, predictive analytics, Big Data and data science.

Based on the issues mentioned, it is clear that the intent of the current research to utilise Big Data as well as data science to intelligently predict insurance claims fraud could add value.

The purpose of this research was to investigate an intelligent way of predicting insurance claims fraud. The discussions in this chapter have shown that the introduction of an intelligent predictor of insurance claims fraud could remedy some issues in the insurance industry. It is therefore important to gain a better understanding of insurance fraud so that this can be achieved. Such investigation will be performed in the next chapter.

3 Fraud and Fraud Prediction in The Insurance Industry

Chapter 2 described the insurance industry and issues that have emerged in it. It related these issues to data and showed how data was commonly used to remedy the issues in the insurance industry. The chapter continued to describe the sources of data in the insurance industry and identified a sub-set of this data, namely claims data.

As was mentioned earlier, one of the main issues currently plaguing the insurance industry is fraud. Fraud was found to be part of both regulatory and insurance modernisation issues. Chapter 2 mentioned the importance of proper data use in the alleviation of fraud. To determine whether data is valuable as part of an intelligent method of predicting fraud, it is necessary to first understand such fraud and to explore fraud in the insurance industry in particular. Chapter 3 describes fraud in the insurance industry as well as fraud prediction in financial systems. It is important to gain an understanding of fraud prediction in financial systems since an insurance claims system is a type of financial system – hence intelligent fraud prediction in other financial systems could be applied to insurance claims systems. It is for this reason that this chapter discusses both financial systems and insurance claims systems and focuses on fraud prediction in these systems. There is literature that relates to intelligent fraud prediction in insurance claims systems (Sharma and Panigrahi, 2013; Hassan and Abraham, 2016), but it would be a shortcoming of this research not to consider research into other financial systems, as well as the work done by Dal Pozzolo et al. (2014) and Zareapoor and Shamsolmoali (2015).

The chapter begins by describing what insurance fraud and insurance claims fraud are and formulates definitions for both terms. Subsequent to this, it describes the possible types of insurance claims fraud as well as a number of examples, and compares the similarities in these examples. The chapter also describes the insurance process so that the occurrence of insurance claims fraud can be better understood. The chapter then moves on to discuss traditional fraud detection techniques and proceeds to discuss existing research involved in predicting fraud in financial systems. This is further expanded based on the research into the use of intelligent fraud prediction techniques in financial systems. The use of such techniques in a sub-set of financial systems – insurance claims systems – is also investigated.

3.1 Fraud

Fraud is considered “an act or instance of deception, an artifice by which the right or interest of another is injured, a dishonest trick or stratagem” (Oxford English Dictionary Online, 2017d), or “criminal deception; the using of false representations to obtain an unjust advantage or to injure the rights or interests of another” (Oxford English Dictionary Online, 2017e). These definitions of fraud are vague and hence need to be narrowed down to apply to the insurance industry.

Insurance fraud can be described as “the wrongful or criminal deception of an insurance company for the purpose of wrongfully receiving compensation or benefits” (Legal Dictionary, 2016), or “criminal acts, provable beyond a reasonable doubt, that violate statutes making the wilful act of obtaining money or value from an insurer under false pretences or material misrepresentations” (Derrig, 2002).

There are clear trends and similarities between these four definitions. One similarity is the use of the terms ‘criminal’ and ‘deception’. Another similarity is the fact that parties are receiving unjust or wrongful benefits through this deception.

Therefore, for the purposes of this research, the four definitions quoted above are used to propose a definition of insurance fraud.

3.1.1 Definition: Insurance Fraud

The criminal deception of an insurance company to gain unjust or wrongful benefits/compensation.

3.1.2 Types of Insurance Fraud

Tennyson (2002) maintains that insurance fraud can be separated into two categories, namely planned fraud and opportunistic fraud. Planned and opportunistic fraud are also known as hard and soft fraud respectively (Goel, 2014). An example of hard fraud could be professionals who contrive false injuries or accidents to gain monetary value. An example of soft fraud, on the other hand, involves individuals over-inflating claims to increase their monetary gain (Tennyson, 2002). Soft fraud can be seen as unethical and abusive behaviour, whereas hard fraud is criminal fraud that can be prosecuted (Tennyson, 2008). According to Minnaar (2000), fraud can only be prosecuted in South Africa if the fraudulent act was unlawful in terms of South African law. Insurance claims fraud need not be prosecuted by the prosecuting authority if they decline to do so; it can be privately prosecuted by insurers (Fulbright, 2014). Therefore, if insurers can predict soft insurance claims fraud and decide to prosecute it, it can become hard fraud.

Wells (2013) of The Association of Certified Fraud Examiners created an Insurance Fraud Classification System that gives an overview of fraud against insurers vs fraud against insured and beneficiaries. Schemes against insurers include false applications, false claims, theft of premiums and receipts, and incorrect payments to beneficiaries (Wells, 2013). Schemes against insured parties include the fraudulent denial of claims, underpayment of claims, internal or third-party theft/ false claims, as well as sales and marketing practices deemed unethical or deceptive (Wells, 2013). This classification system contradicts the definition of insurance fraud that was put forth in Section 3.1.1 and the definitions in the Legal Dictionary (2016) and by Derrig (2002), as they specify the fraud to go against insurance companies. Wells' classification system also falls beyond the scope of this research, but it is important to note that many people view insurance fraud as any fraud involved in the insurance process; not just fraud that affects the insurer.

In line with the reason for this research— fraud costing the insurance industry large amounts of money – the researcher focused his study on schemes against insurers. To limit the research by distinguishing between hard and soft fraud could be detrimental and as such, both shall be included to reduce the limitations of the results.

According to Viaene and Dedene (2004), insurance claims fraud is one of the most common insurance fraud types, but there is also another type of insurance fraud, namely underwriting fraud (Viaene and Dedene, 2004). Thus, the definition of insurance fraud had to be refined to meet the scope of this research, which focuses on insurance claims fraud. Insurance claims fraud is next described below.

3.1.3 Insurance Claims Fraud

The definition of insurance fraud was further refined to fit the scope of this research. This was done by applying the definition of an insurance claim to the definition of insurance fraud in Section 3.1.1. A clear definition of insurance claims fraud could not be derived from existing definitions as credible dictionaries such as Oxford English Dictionary Online (2017k) and Cambridge English Dictionary (2017a) did not contain such definitions at the time of this research. Other definitions of insurance claims fraud were found, but these were not from credible sources.

An insurance claim is “a request to an insurance company for payment relating to an accident, illness, damage to property, etc.” (Cambridge English Dictionary, 2017b) or “a formal request to an insurance company asking for a payment based on the terms of the insurance policy” (Investopedia, 2017).

A common pattern that emerged from the two definitions above was that they were a “request” for “payment” from an “insurance company”. Based on them, an insurance claim was formulated as follows for the purposes of the research at hand:

3.1.3.1 Definition: Insurance Claim

A formal request to an insurance company for payment relating to a loss underwritten in an insurance policy.

This definition of claims was subsequently applied to the definition of insurance fraud and hence, a definition of insurance claims fraud was derived. Insurance claims fraud was described as follows for the purposes of this research:

3.1.3.2 Definition: Insurance Claims Fraud

The criminal deception of an insurance company to gain unjust payment relating to a loss underwritten in an insurance policy.

3.1.3.3 Examples of Insurance Claims Fraud

There have been countless cases of insurance claims fraud, some of which are described below. This is by no means an exhaustive list, but value will still be gained by exploring a few examples. From the many examples of insurance claims fraud, five were chosen owing to the fact that they fall within the scope of this research. The latter was limited to property and casualty insurance and does not include life or medical insurance. The most famous examples of insurance fraud involved life insurance fraud and medical insurance fraud (Brown, 2011; CAIF, 2016) and hence were excluded from the current study. The remaining examples included claims that were facilitated by intermediated insurance companies and the first two of the five examples below were selected because they had occurred in South Africa.

1. **Changing bank details before paying claim:** In one instance an agent from an insurance brokerage submitted fraudulent claims without the policyholders' knowledge. The agent submitted claims on behalf of policyholders. Prior to a claim being paid, the agent changed the account details of the policyholder on the system to his own banking details. Once the claim had been paid, the suspect changed these details back (SAICB, 2014).
2. **Fabricating motor-vehicle claims:** A syndicate was involved in purchasing salvaged high-valued motor-vehicles. Once the vehicle had successfully been financed, it was reported as hi-jacked or stolen and the syndicate submitted an insurance claim. The insurer then settled the value of the motor-vehicle with the finance house (SAICB, 2013).
3. **Committing arson to gain a pay-out:** Once-off insurance claims fraud was also committed by policyholders, for example a Chicago executive who burnt

down his house to claim \$730,000. This offence was exacerbated by the fact that the criminal tried to make the incident look like a suicide by leaving his 90-year-old mother in the building (Macdonald, 2017).

4. **Adding beneficiaries to a claim:** A case of insurance claims fraud was created when a suspect had access to insurance claims and on submission of the claims, false information and beneficiaries were added to the claim. This resulted in the beneficiaries receiving pay-outs when the policyholder received claimed pay-outs. The beneficiaries subsequently paid the money to the suspect (TDI, 2016).
5. **Staging motor-vehicle accidents:** A third-party service provider – a motor-vehicle repair centre – staged motor-vehicle accidents involving deer parts and vehicles that the suspect dismantled to submit insurance claims. This suspect involved loss adjusters, motor-vehicle tow-truck drivers, as well as employees at his repair centre to facilitate the scams (Quiggle, 2017).

From this list of examples, similarities were found that are analysed in Table 3.1. The table lists and compares the different examples of fraud and indicates whether each example contains a similar feature as the other examples. The similarities were decided upon based on their relation to this research. In Section 3.1.2 it was indicated that fraud can be either “*Hard fraud*” or “*Soft fraud*” and hence both types were included in Table 3.1. It was also stated in Chapter 1 that fraud regularly originates from brokers or agents and hence the origin of the fraud was compared (“*Fraud originates from policyholder*”, “*Fraud originates internally from a broker/ service provider*”). It was also stipulated that data mining was used as an intelligent manner of detecting fraud by determining patterns in the data. It is therefore interesting to note whether the perpetrator committed multiple cases of fraud to determine whether “*Patterns*” could be derived.

To understand the table, the first example of insurance claims fraud is explained as follows:

An agent from an insurance brokerage was submitting fraudulent claims without policyholders’ knowledge. Thus, the fraud did not “*originate from policyholder*”, but the “*internally from a broker/ service provider*”. Since the fraud was planned by an individual and the claim was completely fabricated, it can be seen as “*Hard fraud*”. This agent submitted multiple cases of fraudulent claims and hence the fraud was not “*Once-off*” and “*Patterns [could] be derived*”. Based on this example of insurance claims fraud, the reader should be able to understand Table 3.1 and manage to interpret the other examples of insurance claims fraud.

Table 3.1: Similarities and Differences between the Examples of Insurance Claims Fraud

Example	Similarities					
	Fraud originates from policyholder	Fraud originates internally from a broker/service provider	Hard fraud	Soft fraud	Once-off	Patterns can be derived
Changing bank details before paying out claim		X	X			X
Fabricating motor-vehicle claims	X		X			X
Committing arson to gain a pay-out	X		X		X	
Adding beneficiaries to a claim		X	X			X
Staging motor-vehicle accidents		X	X			X

It can be noted from this comparison that all the examples shown were hard fraud. Because soft fraud is more difficult to detect and reduce, it is not documented as frequently (Bhowmik, 2008). The examples also show that fraud often originates internally from a broker or service provider. This reaffirms the statement by Ho (2014) that brokers or agents are commonly involved in insurance claims fraud. What is interesting to note is that when a syndicate is involved (i.e. it is not a single policyholder claiming), there are multiple cases of fraud. Therefore, when fraud has been committed once, it will most likely happen again and patterns can be determined from the data. Such patterns are part of the field of data science, which involves the extraction of knowledge from data patterns (Dhar, 2013). An example of a pattern that can be derived from the second example is the fact that claims were every time submitted directly after the vehicle had been financed. Claims should perhaps be more suspicious if they are submitted soon after a policy has been taken out. Using data science could therefore be helpful to predict insurance claims fraud and to determine when claims fraud could occur in insurance processes.

3.1.4 Claims Fraud's Occurrence in Insurance Processes

To gain a better understanding of when insurance claims fraud may occur, it would be valuable to better understand some of the processes in insurance. Insurance can be

separated into six processes, but these are not limited to product development, marketing, underwriting, claims, administration and investment (Saporito, 2015; Catlin et al., 2015). Figure 3.1 explains these processes by listing its relevant sub-processes. This diagram was derived from research performed by Saporito (2015) and Catlin et al. (2015) and the sub-processes were supplemented with research performed by Saporito (2015); Mahlow et al. (2016); Pulizzi and Heandley (2014); Lee et al. (2007); Rudolph and MAAA (2011) as well as John (1993). The diagram shows each insurance process on the left with its relevant possible sub-processes on the right. The claims process is shown in red so that the occurrence of insurance claims fraud is more visible. The diagram's main purpose is to help the reader understand whether an instance of insurance fraud can be classified as insurance claims fraud or not, for example policyholders misrepresenting themselves on applications such as incorrectly describing the identity of the main car driver (Krawczyk, 2009). The latter can be seen as underwriting fraud and not insurance claims fraud.

The key research question posed in Section 1.4 is “*What elements should a solution have so as to be utilised to intelligently predict insurance claims fraud?*” Based on this question, “*Developing an intelligent manner of predicting insurance claims fraud*” and “*Predicting insurance claims fraud*” was added to the diagram, as they are processes required in this research. They are two separate processes, as a solution that intelligently predicts insurance claims fraud could not be generated at the same time as the prediction of insurance claims fraud. The diagram was not developed into a sequence diagram as the insurance processes do not flow into each other – instead, they often occur concurrently. For example, many policyholders claim while their premium is being invested and hence investment does not happen before or after a claim – it is rather seen as a continuous process.

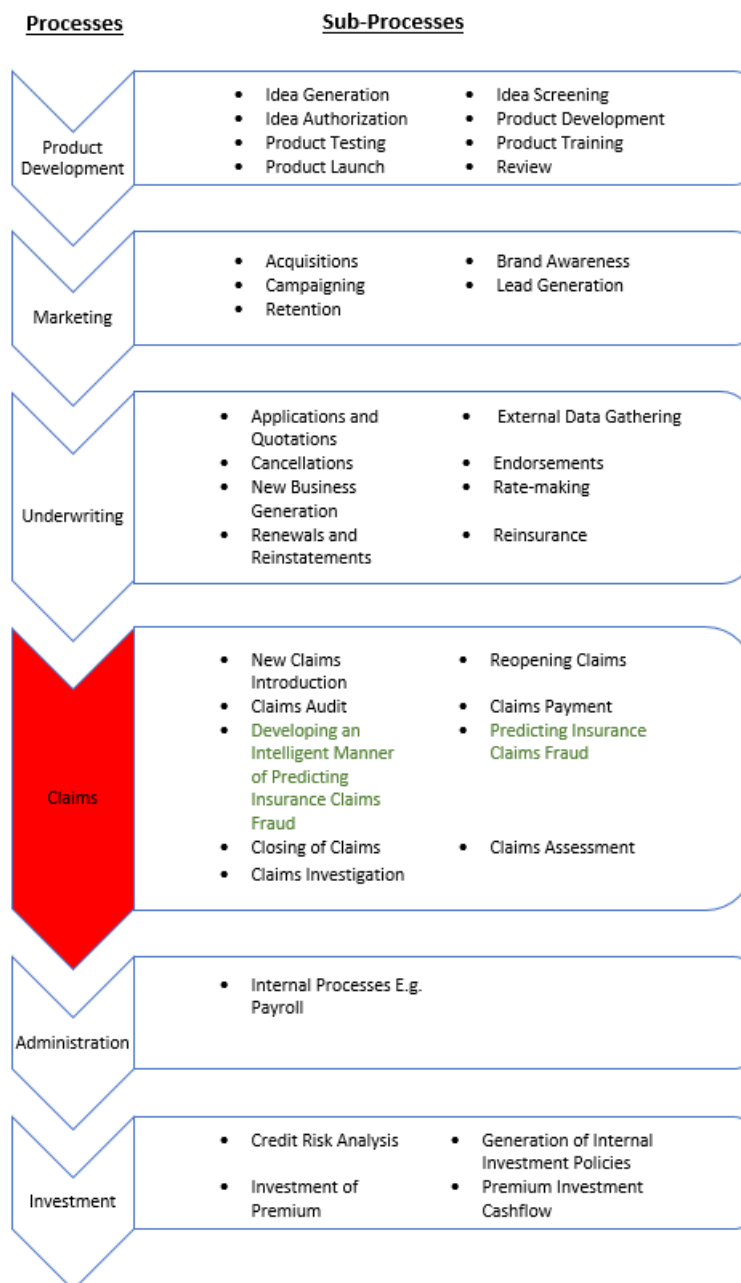


Figure 3.1: Insurance Processes with Sub-processes derived from Saporito (2015) and Catlin et al. (2015)

When viewing this list of processes, it is important to note when the prediction of insurance claims fraud would occur. Since each example in the non-exhaustive list of insurance claims fraud (see Table 3.1) is not a different type of fraud such as underwriting fraud, the prediction of insurance claims fraud should occur during the claims process (marked in red) – perhaps already before claims payment. If the fraud occurs for example during the “Applications and Quotations” sub-process by being deceitful on an application form, then the fraud can be categorised as underwriting fraud and not insurance claims fraud. To accurately identify the exact moment that the prediction of insurance claims fraud should occur, the claims process would need to be examined to arrive at a specific use case and high-level design. This is performed below.

3.1.4.1 The Insurance Claims Handling Process

The insurance claims handling process can be divided into eight generic stages derived from literature contributions by van Jaarsveld et al. (2015) and Olalekan Yusuf and Ajemunigbohun (2015). Olalekan Yusuf and Ajemunigbohun (2015) mention the following nine stages: acknowledging of claims; assigning of claims; identifying the policy relating to the claim; contacting the policyholder; investigating the claim; documenting the claim; determining the cause; determining loss amount; concluding the claim. van Jaarsveld et al. (2015) in turn identify the following seven stages in processing insurance claims: the verbal reporting of claims; claims form completion; assigning of claims to a loss adjuster; processing of claims; assessing of claims; settlement of claims; claims dispute arbitration. Since the literature by both van Jaarsveld et al. (2015) and Olalekan Yusuf and Ajemunigbohun (2015) include valuable processes, it would be a disservice to this research to not include all of these stages. Therefore, the researcher compiled a list containing eight of these process stages, namely: opening of new claims; assigning of claims; processing of claims; assessment of claims; claims investigation; settlement of claims; arbitration of claims; claims closure. These eight stages include what is required by van Jaarsveld et al. (2015) and Olalekan Yusuf and Ajemunigbohun (2015) and are described below based on their research.

1. Opening of New Claims

The claims process starts with a policyholder verbally reporting a claim to an insurance service consultant. The policyholder next reports the claim in writing by submitting the completed claims form and supporting documents within a specified amount of time. If the claim is not reported within an insurer's specified response time threshold, the insurer may reject the claim.

2. Assigning of Claims

The second part of the process involves the assigning of claims to a loss adjuster. Loss adjusters can be internal to (work for) the insurance company or their tasks can be outsourced.

3. Processing of Claims

The loss adjuster and insurer parties then process the claim details, such as details of the event and details of the people involved in the event. The insurer and loss adjuster also analyse previous claims by the claimant.

4. Assessment of Claims

The adjuster determines the existence of an insurance policy relating to the claim and validates that all premiums have been fully paid. The adjuster also determines the coverage of the policy and whether the claim is covered by the policy. Once this is done, the adjuster determines the loss and the remuneration in terms of the claim.

5. Claims Investigation

If the adjuster finds that the claim is suspicious, the claim is submitted to an internal or external investigator who will then investigate the matter more closely.

6. Settlement of Claims

If the investigation does not reveal anything wrong, the claim can be settled. The policyholder is informed of the result of the claims assessment and of the remuneration/repair or reinstatement of the loss.

7. Arbitration of Claims

If a client is not satisfied with the result of the claims process, the insurer will appoint an arbitrator to settle a dispute.

8. Claims Closure

Once all parties are content with the result of a claim, the claim is closed.

These eight generic stages occurring during the insurance claims handling process give an idea of how claims are received, evaluated and paid. The different processes can be depicted by a process flow, and Figure 3.2 provides an illustration with the actors and systems involved in each stage. The diagram shows each phase and indicates the corresponding sub-section number on the left-hand side. It also shows that there are four main actors in the claims-handling process, namely the policyholder, a service consultant, a loss adjuster and a claims investigator. One main system is involved in the process, namely the claims administration system.

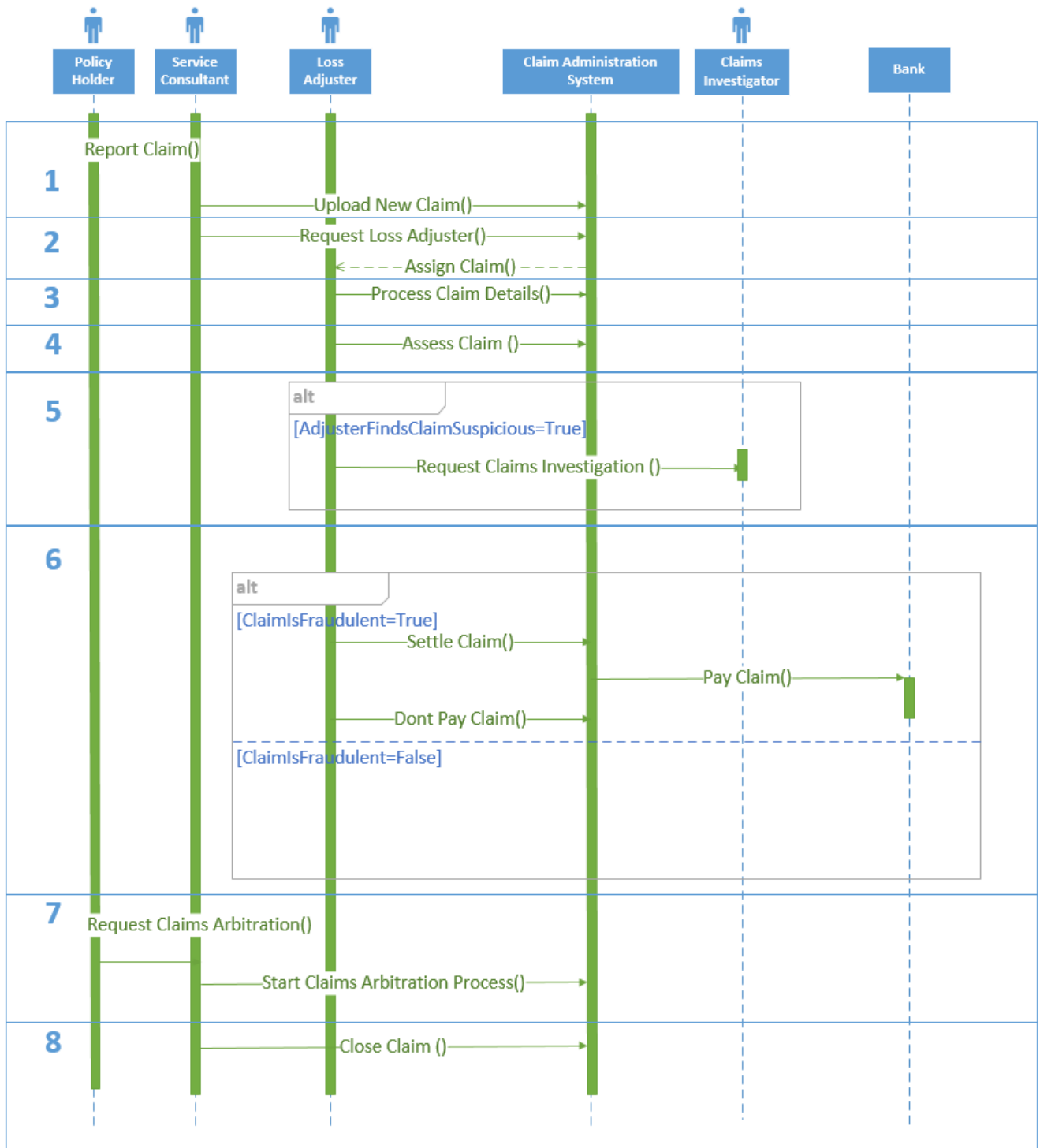


Figure 3.2: The Insurance Claims Handling Process derived from van Jaarsveld et al. (2015) and Olalekan Yusuf and Ajemunigbohun (2015)

Because the purpose of this research is to investigate an intelligent method of predicting insurance claims fraud, enriching these phases to allow for the prediction of insurance claims fraud is necessary. Exactly when this needs to be done, will be explored in the next section.

3.1.4.2 Adding Fraud Prediction to the Insurance Claims Process

Various examples of insurance claims fraud have been put forward in this research (see Section 3.1.3.3). The researcher revisits these examples to try and establish when insurance fraud occurs and would need to be picked up. Table 3.2 contains a basic summary of fraudulent activities and suggests the last stage of insurance claims handling where the described fraudulent activity could possibly have occurred.

Summary of Example	Last possible occurrence in claims process
1. Claim submitted by broker without policyholder being aware of it (SAICB, 2014).	Assessment of claims
2. Syndicate purchasing high-valued motor vehicles, financing them, insuring them, reporting them as stolen and keeping the vehicle (SAICB, 2013).	Opening of new claims
3. A policyholder committing arson by burning down his own house to claim \$730,000 (Macdonald, 2017).	Opening of new claims
4. A broker adding false beneficiaries to claims to gain claims pay-outs (TDI, 2016).	Processing of claims
5. A motor vehicle repair centre staging accidents to claim money from insurers (Quiggle, 2017).	Assessment of claims

Table 3.2: The Occurrence of Fraud in the Claims-handling Process

From this table it is clear that the insurance claims handling process requires an additional process occurring between the assessment of claims and claims investigation that would flag possibly fraudulent claims. This process could not be introduced at an earlier stage, as the examples show that the fraud could be committed at any time up until investigation. Therefore, an intelligent method of determining and predicting insurance claims fraud should be incorporated into the claims-handling process right before the investigation of a claim. Both traditional fraud detection techniques and intelligent fraud prediction is discussed below.

3.2 Traditional Fraud Detection Techniques

For the purposes of this research, a traditional fraud detection technique would include fraud detection techniques that can be described as long established (Oxford English Dictionary Online, 2017j). In terms of these long-established techniques, fraud needs

to be picked up by external auditors, accountants or internal whistle-blowers (Singleton et al., 2006). The research by Othman et al. (2015) agrees with this statement and mentions the fact that internal and external audits, whistle-blowing hotlines, policies, penalties and disciplinary action help to prevent fraud. From the examples mentioned by Othman et al. (2015), fraud detection can be performed only through auditing and whistle-blowing. Policies, penalties and disciplinary action could well prevent fraud, but they do not necessarily detect or predict it. Based on the research mentioned, fraud auditing, forensic accounting and whistle-blowing can be isolated as traditional fraud detection techniques that can add value in the detection and prediction of fraud.

3.2.1 Fraud Auditing

Auditing can be considered a fraud detection process. For fraud prediction through auditing, insurance companies would use a “Red Flag” process. This means that if a claim was given certain flags to indicate that it was suspicious, the claim would be further investigated by an insurance claims investigator (Dionne et al., 2003). Since these red flags could be used by criminals to defraud insurance companies, they are not commonly published (Dionne et al., 2003). According to Grabosky and Duffield (2001), a red flag is a variation or anomaly from normal or predictable patterns, such as a sudden change in activity. The red flag process involves viewing a claim, determining whether the claim contains a red flag or a combination of red flags and subsequently investigating the claim (Insurance Fraud Working Group, 2011). Examples of red flags include policyholders that are verbally aggressive when applying for insurance policies, policyholders insisting on cash payments, and policyholders giving evasive answers (among others) (Insurance Fraud Working Group, 2011). Auditing to detect insurance claims fraud is problematic as auditors often use sampling during the audit process and hence fraudulent transactions can easily be missed (Morton, 1993).

What can be learnt from fraud auditing for the purposes of this research, namely creating an intelligent manner of predicting insurance claims fraud, is the fact that the “Red Flag” process could be used to predict fraud. If a claim is processed and a system flags the claim as potentially fraudulent, then this could be seen as an intelligent manner of predicting fraud. If normal patterns could be derived from data, then anomalies or deviations from these patterns could be flagged in a system in order to intelligently predict fraud.

3.2.2 Forensic Accounting

Forensic accounting is a field that includes accounting, auditing and investigative skills to determine fraud that is suitable for use as evidence in court (Bhasin, 2007). Forensic accounting would be used for litigation support when fraud has been detected

(Kranacher et al., 2010). Insurance companies use forensic accountants to accurately assess whether claims should be paid out and what the settlement should be (Bhasin, 2007). The forensic accountant supplements the financial knowledge regarding an insurance claim with further information (Mohn, 2013) by working with the policyholder and loss adjuster to assess the loss and determine all of the possible factors that will affect the payment of a claim. This is considered problematic, as fraud would need to be detected before it could be investigated by a forensic accountant.

For the purposes of the current research, forensic accounting could be supplemented with evidence provided by an intelligent system that predicts fraud. After the “Red Flag” process has been performed and a claim has been investigated, fraud would still need to be proved. If a system that intelligently predicts fraud is able to supply a forensic auditor with necessary evidence, such as a list of suspicious activities that is required to prove that the claim is fraudulent, such system could be seen to supplement the forensic accountant’s work.

3.2.3 Whistle-blowing

Whistle-blowing involves a questionable event that is noticed by an employee who evaluates this event and determines whether to “blow the whistle on the event” (Dworkin and Baucus, 1998). According to Todd et al. (1999), whistle-blowing is less prevalent in the insurance industry than in other industries, which means that it could be seen as less effective in this context.

The offence reported by the whistle-blower would still need to be proved by a fraud investigator. If a system that intelligently predicts fraud could supply the investigator with evidence of suspicious activity, it could aid the success of the whistle-blowing process.

The above discussion of three traditional fraud detection techniques (fraud auditing, forensic accounting and whistle-blowing), indicates that traditional fraud detection could be improved upon. Intelligent fraud prediction that can predict fraud in advance, rather than detect it long after a claim has been processed, can add great value in the insurance industry. This technique is described below.

3.3 Intelligent Fraud Prediction

Owing to the fact that the purpose of this research was to investigate an intelligent method of predicting insurance claims fraud, it was necessary to formulate a definition of intelligent fraud prediction. Such a definition could not be derived from existing definitions, as credible dictionaries such as Oxford English Dictionary Online (2017k) and Cambridge English Dictionary (2017a) did not yet contain such definitions at

the time of this research. Moreover, definitions that did exist were not from credible sources. A definition of intelligent fraud prediction therefore had to be compiled from the definitions of “intelligent”, “fraud” and “prediction”.

The term “intelligent” (or “intelligence”) is defined as “having a high degree or good measure of understanding” (Oxford English Dictionary Online, 2017f). This intelligence is broad hence intelligence with regard to computers and systems could be more valuable. In terms of this research, “intelligence” can be seen as “the ability for an information processing system to adapt to its environment with insufficient knowledge and resources” (Wang, 2007). According to Beni (2004), there are no definitions of intelligence that can satisfy every use case, but the researcher believes that intelligence is behaviour that is neither predictable nor random. Although there are contrasting opinions of what intelligence is, the definitions above can add to the definition of intelligent fraud prediction formulated in this research. There is not much overlap between these definitions, so the approach taken was to take those parts of each definition that applied to this research. The following parts could add value: intelligence needs understanding; it functions with insufficient knowledge; and the system needs to adapt to its environment.

It was previously specified that fraud is “an act or instance of deception, an artifice by which the right or interest of another is injured, a dishonest trick or stratagem” (Oxford English Dictionary Online, 2017d), or “criminal deception; the using of false representations to obtain an unjust advantage or to injure the rights or interests of another” (Oxford English Dictionary Online, 2017e). These two definitions illustrate that deception and injuring the rights of another are key elements of fraud.

The definition of prediction is “the action of predicting future events; an instance of this, a prophecy, a forecast” (Oxford English Dictionary Online, 2017h), or “a statement about what you think will happen in the future” (Cambridge English Dictionary, 2017c). The key pattern in these two definitions is that the action will take place in the future and it is determining what will happen or the events that will happen at this point.

To summarise, the definition of intelligent fraud prediction can be formulated as follows.

3.3.1 Definition: Intelligent Fraud Prediction

Intelligent fraud prediction involves using a system that can adapt to the environment by using understanding, but that lacks the knowledge to pre-emptively determine whether a person has gained an unjust advantage over another.

According to West and Bhattacharya (2016), intelligently predicting and detecting financial fraud is generally applied to three areas, namely bank fraud, corporate fraud

and insurance fraud. Because this research focuses its efforts on predicting such fraud, the researcher proceeds to discuss associated literature below. The remainder of this discussion is devoted to financial systems and insurance claims systems respectively. Although an insurance claims system is a type of financial system, greater emphasis needs to be placed on insurance claims systems as they constitute the focus of this research.

3.3.2 Intelligent Fraud Prediction: Financial Systems

The following section gives a brief description of intelligent fraud prediction in financial systems by exploring some of the existing literature. It subsequently describes two literature contributions that could add insight when applied to this research.

The advent of Big Data has resulted in its utilisation together with data science to predict fraud in financial systems (Yoon et al., 2015). Not only was Big Data invoked to predict fraud in financial systems, it was also used during the audit process to detect past fraud (Yoon et al., 2015). Much literature has been published about fraud detection techniques through applying data science in financial systems, as indicated in the work of Ngai et al. (2011). A large proportion of this literature focuses on efforts to detect credit card fraud (Dal Pozzolo et al., 2014; Zareapoor and Shamsolmoali, 2015) and online banking fraud (Wei et al., 2013; Carminati et al., 2015) – both of which can be seen as examples of bank fraud. Within bank fraud, the methods that are discussed involve using supervised and unsupervised machine learning and data-mining techniques.

3.3.2.1 Brief Overview of Machine Learning as a Fraud Prediction Technique in Financial Systems

The supervised machine learning techniques commonly include tree-based and rule-based algorithms and neural networks (Khan et al., 2014; Behera and Panigrahi, 2017; Sahin et al., 2013). Tree-based methods use sets of predictor variables and generate trees that classify a response variable based on these sets of predictor variables (Bell, 1999). Rule-based methods are algorithms that use data or specified rules that have been derived from human expert knowledge to solve problems (Abraham, 2005). Neural networks are algorithms that are based on the neuro-anatomy of animals and humans. They are derived of a set of weighted connections that take inputs and yield an activation as its output (Gallant, 1993). These supervised machine learning techniques take data sets that have been classified into fraudulent and non-fraudulent classes to generate a model and subsequently classify new transactions into these classes based on the model. This matches this research's definition of intelligent fraud prediction, as each of these methods takes data as an input and adapts to its environment by attempting to understand the data. These techniques subsequently use this understanding

to determine what will happen in the future and to establish whether deception has occurred.

The unsupervised machine learning techniques include clustering algorithms such as self-organising maps and K-nearest neighbours (Ngai et al., 2011). Self-organising maps are algorithms that convert the relationships of high dimensional data into simple geometric relationships so that the statistical relationship between complex data sets can be understood (Kohonen, 1998). K-nearest neighbours is a classification algorithm that assigns unclassified transactions into groups based on their distance from “k” neighbours (Denoeux, 1995). These unsupervised machine learning techniques are primarily used to detect whether transactions do not fall within normal groups and as such must be flagged as anomalies. Once they have been flagged as anomalies, the transactions can be investigated for fraud (Ngai et al., 2011). Again, these techniques can be seen to match the current research’s definition of intelligent fraud prediction as they take data as an input and use this data to adapt to their environment by attempting to gain understanding from the data. These techniques subsequently use this understanding to determine whether new transactions fall into normal groups and if they do not, they are predicted to be fraudulent, as deception has probably occurred.

The aforementioned data science methods are currently supplemented with Big Data as well as analytics, and they are researched in the banking sector to analyse past transactional trends. According to Srivastava and Gopalkrishnan (2015), this technique can be used in conjunction with behavioural analysis to indicate potential threats. Although there is a high chance of creating false positives, this can be an effective method of predicting transactional and banking fraud if it is used as part of a holistic approach (Srivastava and Gopalkrishnan, 2015).

The above is a brief overview of the research that has been performed to predict fraud in banking systems, but it does not yield enough information about how this has been accomplished and how accurate it was. The researcher therefore examines two of these literature contributions so that a better understanding of the application of intelligent fraud prediction can be gained. Of the mentioned research into online banking fraud and credit card fraud prediction, Dal Pozzolo et al. (2014) and Wei et al. (2013) are more widely referenced than Sahin et al. (2013), Khan et al. (2014), Zareapoor and Shamsolmoali (2015) and Behera and Panigrahi (2017). The research of Ngai et al. (2011) is most widely referenced but it was already six years old at the time of this research. Hence further description of the research by Dal Pozzolo et al. (2014) and Wei et al. (2013) could add value.

3.3.2.2 In-depth Examples of Machine Learning as a Fraud Prediction Technique in Financial Systems

Dal Pozzolo et al. (2014) compare machine learning techniques that can be applied

to large data sets to detect fraud. Their research mentions both supervised and unsupervised machine learning techniques but focuses on supervised machine learning techniques. Past transactional trends and normal behaviour are analysed so that when a new credit card transaction comes in, it can be evaluated as fraud or not fraud. According to Dal Pozzolo et al. (2014), each fraudulent transaction must be treated equally, as although a low-value transaction may not affect a credit card company as much as a high-value transaction, the low-value transaction is often a “test” from the fraudster and a large transaction may be expected to come next. What is interesting to note is that if the supervised machine learning technique predicts that there is a high probability of fraud for a transaction, a credit card is automatically blocked and if the probability is over a fraud risk, then an investigator investigates the transaction. To test the different supervised machine learning techniques, the research (Dal Pozzolo et al., 2014) used a legitimate dataset from a payment service provider in Belgium and generated additional aggregated variables for the data set. The fields included information such as the currency, the transaction amount, the point of sale and merchant information. The data set, which is unfortunately not available owing to privacy considerations, trained Support Vector Machines, random forests and neural networks after balancing the data (50% fraudulent, 50% non-fraudulent). Models created were determined to be more accurate when retrained weekly or every 15 days, and they had larger data sets. The research by Dal Pozzolo et al. (2014) mentioned that Random Forests outperformed the other algorithms but did not give a percentage accuracy. Unfortunately, the research did not put forth any example rules or trees, but this could be because the research did not focus specifically on a tree- or rule-based algorithm.

It is our opinion that this research should rather be seen as prediction as – although the researchers referred to the application as detection – suspicious credit card transactions that came in were flagged and hence “predicted” before an audit process. The research did not stipulate why only supervised machine learning algorithms had been used and hence this could be seen as a shortcoming. The researchers however mentioned that they intended to perform the research with algorithms that could handle unbalanced data, as most data sets that included fraudulent transactions were unbalanced.

In contrast to the literature by Dal Pozzolo et al. (2014), Wei et al. (2013) proposed an online banking fraud detection system. This system included three machine learning techniques, namely neural networks, contrast pattern mining (Bay and Pazzani, 2001) and decision forests, which were supplemented with domain knowledge from experts. These techniques intelligently detected whether new online transactions were fraudulent by passing the transaction through a system that subsequently scored the transaction based on its risk of being fraudulent. The literature (Wei et al., 2013) states that this technique was effective for predicting fraud when it used data sets that were unbalanced. It is also stipulated that as the sample size of the data that trained the system increased, the accuracy also increased. Seeing that sample sizes could get extremely large, it was

necessary to extract only the relevant variables when training the system. Important characteristics of online banking fraud included the fact that it was dynamic, fraud was hidden when behaviour varied a lot, and fraudsters were intelligent and constantly adapted to new fraud detection techniques. A tool that detects fraud also needed to be instantaneous, as it might be too late if the fraud was detected after the fact. An interesting rule to note was that if transactions were completed in less than three seconds after login, they had a higher chance of being fraudulent – this was too fast for a human user and likely originated from a bot or machine. Another rule to note was that if a user did not view the home page of a website first and did not print their payment confirmation, this constituted suspicious behaviour.

Unfortunately, Wei et al. (2013) did not provide specific technical detail such as the number of layers in the network, but since the research used three machine learning techniques instead of just one, such technical detail could make the literature too cumbersome. The researchers stipulated that it was important to use more than one machine learning technique in a system that intelligently predicts fraud as the advantages of each technique were used to make the system more accurate. This use of multiple machine learning algorithms in conjunction with domain knowledge could have a detection rate of up to 67%.

The research by Wei et al. (2013) could also be seen as online banking fraud prediction rather than detection, as the transactions would be identified as fraud in near-realtime. Although they were specific to online banking fraud, the findings by Wei et al. (2013) could be applied to the current research, as the methods used to increase the accuracy would also add value to insurance claims fraud. Having more than one machine learning algorithm in a system that intelligently predicts insurance claims fraud could increase the accuracy and hence should be used.

3.3.3 Intelligent Fraud Prediction: Insurance Claims Systems

This section describes intelligent fraud prediction in insurance claims systems. Because a key research question (sub-question 2 in Section 1.4) was “*Can new developments in Big Data as well as in data science help to predict insurance claims fraud?*”, it is necessary to note whether data science has been supplemented with Big Data in an intelligent manner to predict insurance claims fraud. The literature contributions regarding the use of Big Data and data science are often not related to property and casualty insurance, such as in the case of medical insurance (Wang et al., 2017) or they are focused on a specific sub-set of this insurance, for example automobile insurance (Li et al., 2016). This is not part of this research and would constrict the scope of this research.

Past literature contributions that incorporated Big Data into data science to predict insurance fraud were often theoretical and had not been tested in a practical application

(Power and Power, 2015). The literature was also not focused on short-term insurance, instead it related to medical insurance (Shi et al., 2016; Raghupathi and Raghupathi, 2014). Since machine learning is commonly used in intelligent fraud prediction, the available and existing literature is described below. This section continues to describe two of the relevant literature contributions.

3.3.3.1 Brief Overview of Machine Learning as a Fraud Prediction Technique in Insurance Claims Systems

Research has been performed into the use of data science and machine learning techniques as intelligent methods to predict insurance claims fraud. These techniques include neural networks (Xu et al., 2011), decision trees (Hassan and Abraham, 2016), logistic regression (Sharma and Panigrahi, 2013) and Bayesian networks (Bhowmik, 2011). These techniques (as used in the works referenced) are supervised machine learning techniques. Decision trees are tree-based classification algorithms that can classify transactions into classes. This happens when the transaction passes through decisions in the tree and reaches the final node, which can be seen as its class (Westreich et al., 2010). Hassan and Abraham (2016) generate a decision tree and train the tree with unbalanced claims data by sampling the claims data and generating a new dataset with data that is more balanced. This is subsequently used to classify claims into fraudulent and nonfraudulent classes. Logistic regression takes a set of continuous input variables and predicts a binomial output value with a confidence (Sperandei, 2014). In the case of insurance claims fraud, variables of insurance claims are passed into a regression model and a binary indicator determines whether they are fraudulent or not. Lastly, Bayesian Networks are used as a graphical method to display conditional interdependencies between variables (nodes) in a network (Ghahramani, 2001). Bhowmik (2011) utilises variables such as the rating of an insured driver (i.e. whether they are the policyholder and whether a report was filed) to classify claims into both legal vs non-legal classes and fraudulent vs non-fraudulent classes.

These four supervised machine learning techniques also comply with the definition of intelligent fraud prediction. This is because these techniques take insurance claims data as an input and adapt to its environment by generating rules such as the following: if the driver is under the age of forty, has a driver rating of one and drives a two-year-old vehicle, then the odds are high that the claim is fraudulent (Bhowmik, 2011). It then uses this understanding to predict whether deception has occurred which will result in a future claim being classified as fraudulent.

Fraud prediction has also occurred by using unsupervised machine learning techniques such as support vector machines (Nian et al., 2016), K-Means clustering (Thomas, 2017) and anomaly detection (Rawte and Anuradha, 2015). The support vector machine algorithm takes in a new claim and determines whether it is normal or not (Kir-

lidog and Asuk, 2012) – in this case, fraudulent or non-fraudulent. K-means clustering segregates a certain number of data records into k-number of clusters (Thomas, 2017). To predict insurance claims fraud, claims that do not fit into these clusters could be flagged as unusual and potentially fraudulent. Lastly, anomaly detection determines the probability of a claim being fraudulent by examining the current claim, compared to previous claims (Rawte and Anuradha, 2015). Regarding these three types of unsupervised machine learning techniques, it is noteworthy that they adhere to our current definition of intelligent fraud prediction. It predicts whether a claim needs to be classified as fraudulent in the future by taking input as data, gaining understanding from that data and determining whether the claims do not fall into normal groups.

To gain a better understanding of the existing literature contributions, the two contributions that relate best to this dissertation are discussed below. Firstly, “A review of financial accounting fraud detection based on data mining techniques” by Sharma and Panigrahi (2013) applies to this research most for a number of reasons: it includes elements of traditional fraud detection methods; it uses data science Such a as an intelligent manner of predicting fraud; it considers Big Data; and it focuses on insurance fraud. Secondly, “Research and application of random forest model in mining automobile insurance fraud” by Li et al. (2016) applies equally most to this research as it includes data science as an intelligent manner of predicting fraud, considers Big Data and focuses on insurance fraud.

3.3.3.2 In-depth Examples of Machine Learning as a Fraud Prediction Technique in Insurance Claims Systems

Sharma and Panigrahi (2013) describe the use of data mining as a data science technique to detect fraud. Their article is a review of existing literature that uses data mining to detect fraud. It does not focus specifically on insurance fraud, but frequently applies the literature to insurance fraud. It describes the fact that data mining can be classified into six classes, namely classification, clustering, prediction, outlier detection, regression and visualisation. The research reviewed supervised and unsupervised machine learning techniques. Although Sharma and Panigrahi did not put forward an opinion on the best technique to predict fraud, they did mention that regression, such as logistic regression and neural networks, are valuable for fraud detection. Both regression and neural networks have advantages and disadvantages when used for fraud detection. Advantages of logistic regression include that it is widely used for fraud detection and also has great explanation ability. This is important for forensic accounting as its purpose is to prove that a claim is fraudulent – an explanation therefore adds great value.

The disadvantages of logistic regression on the other hand include its lack of ability to perform classification – when compared to neural networks. In comparison to re-

gression, the advantages of neural networks are that they do not have stringent data requirements and consequently adjust and generalise well. This serves as a valuable benefit for insurance claims, where the data can be different for various insurers and brokers. The disadvantages of neural networks include their hidden structure, difficulty in explanation and complexity in accuracy. Neural networks' lack of explanation ability is an important shortcoming in respect of insurance claims fraud prediction, because if the fraud needs to be explained by a forensic accountant, the reasoning would need to be given. According to Sharma and Panigrahi (2013), logistic regression can reach up to 95.1% accuracy in predicting fraud, which seems remarkably high. Unfortunately, they did not indicate the level of accuracy of neural networks.

For the purposes of this research, an advantage as well as a downside of the article by Sharma and Panigrahi (2013) is that it is broad and applies to many types of financial accounting fraud. This results in the research not being fully applicable to insurance claims fraud, in other words it is too broad for the scope of the current research. What can be gained from their research, however, is that the accuracy of logistic regression seems to be quite high for a machine learning technique and as such the technique could be a good machine learning technique for predicting insurance claims fraud.

Regarding the second valuable literature contribution, Li et al. (2016) describe the use of random forests as a technique to detect automobile insurance fraud. This was conducted by discussing some past research that included supervised and unsupervised techniques. The Random Forest method combines classification and prediction to determine whether automobile insurance claims are fraudulent. Although the literature (Li et al., 2016) does not specifically reference Big Data, it suggests that the methods used can utilise large data sets, which can be seen to fit the Big Data category. According to Li et al. (2016), an advantage of the Random Forest model is that it can take large numbers of variables and determine the most important variables for predicting fraud. Some of the variables mentioned include the driver's gender, the name of the repair shop, and whether there are damage photos or not. Another advantage of using the Random Forest model that was generated is that it was not affected by outliers and noise.

A disadvantage of the research by Li et al. (2016) is that it is too specific and does not apply to all property/casualty insurance claims. It can be seen however that this research is an intelligent manner of predicting insurance claims fraud as it can use large data-sets. The problem was that this was not proven as sample-sizes were small. Owing to the fact that it has previously been shown that a typical claim would have large amounts of categorical data, the fact that the research had to convert categorical variables to numeric variables could be problematic. Li et al. (2016) valuably shows that pre-processing of data is an important part of generating a model that can intelligently predicting insurance claims fraud.

What is noticeable, is that neither the study by Sharma and Panigrahi (2013) nor

that by Li et al. (2016) included privacy considerations. Although they conducted substantial research with regard to insurance fraud, data science and Big Data, they did not combine these three elements with privacy considerations to create an intelligent way of predicting insurance claims fraud. Thus, the existing research does not apply directly to the property and casualty domain, and privacy requirements of countries have not necessarily been considered.

3.4 Discussion

This chapter looked at fraud in the insurance industry by describing fraud, types of insurance fraud, examples of insurance claims fraud and insurance processes. It also discussed traditional fraud detection techniques as well as intelligent fraud detection techniques in financial systems and insurance claims systems. It appeared that fraud was a substantial problem in the insurance industry and in particular within insurance claims processes. From the research presented, a number of key points were noted:

- The cases of insurance claims fraud that were detected could mostly be classified as hard fraud.
- Patterns could be derived in insurance claims fraud and hence data science can be used to predict these patterns.
- Two separate processes had to be developed: (a) the prediction of insurance claims fraud and (b) the development of a model to predict insurance claims fraud.
- The insurance claims handling process required an additional process that would occur between the assessment of claims and claims investigation to flag possibly fraudulent claims.
- Big Data and data science can be used in financial systems and insurance claims systems to intelligently predict fraud. Big Data is important, as insurance companies have large sets of data. However, data science is equally important as it is often the foundation of intelligent fraud prediction.
- Both supervised and unsupervised machine learning can be used in financial systems and insurance claims systems to intelligently predict fraud. Existing literature shows that logistic regression can be used with a high level of accuracy in intelligent financial systems that are used to predict fraud.
- Combining more than one machine learning technique in a system that intelligently predicts insurance claims fraud can increase its accuracy.
- Models that are used in financial systems to intelligently predict fraud should be regularly trained.

- Existing literature is either too broad or too specific for the scope of the research in hand.
- The existing literature does not focus on privacy.
- Interesting inference can be gained from rules put forth in this chapter. An example of this is the fact that timing is important in online banking; if transactions are too quick, they have a higher likelihood of being fraudulent. Furthermore, rules can depict normal behaviour and deviations from the norm should be flagged as suspicious. For example, if a person does not print a confirmation after performing a transaction, it is suspicious.

It should be clear from the above that predicting insurance claims fraud in an intelligent manner could add much value to the insurance industry. Since this chapter showed that Big Data and data science can be used to intelligently predict fraud, a better understanding of both these concepts needs to be fostered. A theoretical understanding of Big Data Science is required, which is a combination of Big Data and data science. Advantages and disadvantages of data science methods need to be discussed in the following chapter in order to develop a potentially intelligent way of detecting insurance claims fraud. What should be noticed from Chapter 3 is that machine learning is commonly used in intelligent methods of predicting fraud and hence, our discussion of Big Data Science needs to focus on machine learning in Chapter 4.

4 Big Data Science

The previous three chapters regularly mentioned Big Data and data science. This chapter describes the combined term “Big Data Science” by expanding on the principles of Big Data and data science. It was previously determined that Big Data, data science and machine learning are regularly used in intelligent methods of predicting insurance claims fraud. It could therefore add value to gain a theoretical understanding of each of these concepts to comprehend what other techniques involved in them could be used to intelligently predict insurance claims fraud.

This is achieved in Section 4.2 where the three V’s of Big Data are discussed. Next follows a description of Big Data and the tools that can be used to facilitate Big Data. As the purpose of this research is to find an intelligent method of predicting and identifying fraudulent insurance claims, and as it was determined that data science was appropriate for this, a better understanding of data science is also required. Some of the tools used to perform data science are described. Privacy considerations in data science are also covered, as the current research has privacy considerations as a constraint.

4.1 Big Data

Gartner (2016) describes Big Data as “high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation”.

Although this definition of Big Data is accurate and valuable, comparing it to alternate opinions is important. According to Jacobs (2009), Big Data is “data whose size forces us to look beyond the tried-and-true methods that are prevalent at that time”. Thus we need to compare it to a definition that Big Data is “complex, unstructured, or large amounts of data” (Intel IT Center, 2012) or that Big Data “refers to environments in which data sets have grown too large to be handled, managed, stored, retrieved, etc., in an acceptable timeframe” (Slack, 2012).

From the above four definitions, similarities can be derived. What is noteworthy in each definition is the fact that Big Data implies that there are large sets of data that come from multiple sources. It also seems that alternate processing methods are commonly needed for these large sets of data.

Therefore, a new definition of Big Data is formulated as follows:

4.1.1 Definition: Big Data

Big Data is data sets with high volume, velocity and/or variety that result in advanced computing requirements and data-processing methods.

Having large sets of data available does not add much value, unless insight can be gleaned from such data. This is why the term Big Data has morphed into Big Data Analytics/ Big Data Science. Big Data Analytics is a combination of two technologies, namely large sets of data and advanced analytics (Russom et al., 2011). For the purposes of the current research, the terms Big Data Science and Big Data Analytics shall be used interchangeably. The overlap between Big Data and data science can be seen in Figure 4.1.

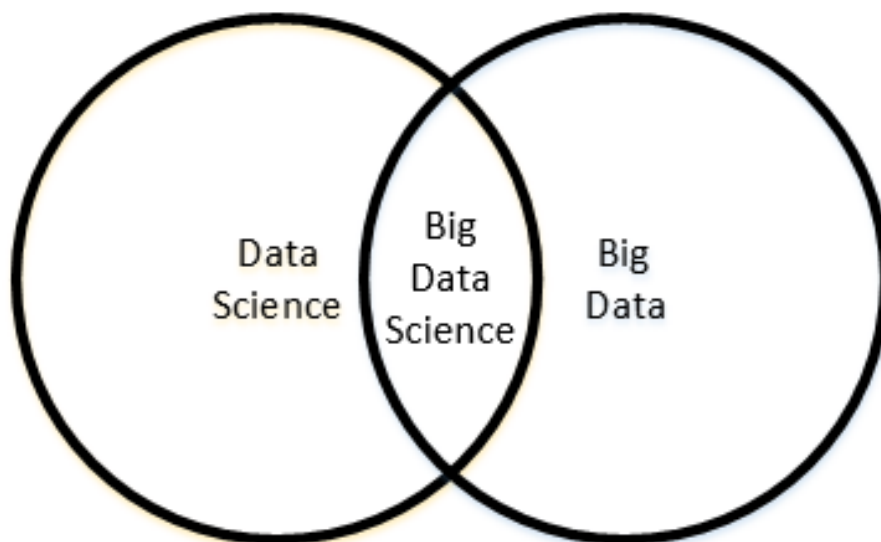


Figure 4.1: The Overlap of Data Science with Big Data

Based on this definition and overlap, the three V's of Big Data, namely Variety, Velocity and Volume, as well as its advanced computing requirements are described below.

4.1.2 Variety

The variety of Big Data results from the fact that the sources of data may be unstructured, semi-structured or structured. The structured data sources can for example be in a tagged, fixed-field data warehouse that is tabled and understandable. In contrast,

unstructured data sources are difficult to understand as the data is not tagged, does not follow a structure or fixed-field format, and as such, deriving meaning from the data is difficult. Lastly, semi-structured data sources do not have fixed fields, but they are understandable as the data is tagged (Sagiroglu and Sinanc, 2013; Zikopoulos et al., 2011).

According to Hussain and Prieto (2016), insurance data originates from a variety of data sources and occurs in the form of both structured and unstructured data. An example of structured data would be the organised transactional data, whereas the unstructured data (which can cause a problem for the insurance industry) is free-text responses regarding claims that are difficult to interpret (Hussain and Prieto, 2016). In the past, insurance companies used only the structured data and not the unstructured data (Mills and Forder, 2012). Utilising the latter may be advantageous to insurers, but it is problematic to determine exactly what unstructured data should be used (Mills and Forder, 2012). Therefore, insurance data can be said to include the first V of Big Data, namely “Variety”.

4.1.3 Velocity

The velocity requirement of Big Data indicates that the processing of the data must be fast. It should occur in a streaming process, due to the vastness of the data that needs to be processed as it enters the organisation. The end user should see the analysis and processing of Big Data as occurring in real-time or near real-time (Sagiroglu and Sinanc, 2013; Zikopoulos et al., 2011).

In the insurance domain, this could include data such as customer feedback data which is continuously streaming into the organisation (Hussain and Prieto, 2016). The velocity can also apply to decision support systems where real-time feedback would need to be given, for instance determining whether an insurance claim should be paid, or whether a policy should be accepted (Hussain and Prieto, 2016). According to Mills and Forder (2012), long-established methods in insurance collate data in batches at certain times before it would be available. This has been changed so that data is readily available and, as such, alternate computing methods need to cater for real-time data access.

4.1.4 Volume

The volume of Big Data is so large that it has changed traditional data-storing methods. Data storage metrics can range from terabytes to petabytes (Sagiroglu and Sinanc, 2013; Zikopoulos et al., 2011).

In the financial services sector, transactional volumes are increasing and in developing countries, the rate of increase is as high as 20% (Hussain and Prieto, 2016). A reason

why the volume of data is increasing in the insurance industry is that regulators are requiring more data from insurers and hence more data needs to be kept. Furthermore, the digitisation of financial services makes it easier for people to transact, and hence the number of transactions is increasing (Hussain and Prieto, 2016). Lastly, data is increasing in the insurance industry owing to the fact that aggregator websites are creating more quotes per insurer than in the past (Mills and Forder, 2012). According to Liao and Zhu (2014), increasing the size of a training set of a model can increase its accuracy and hence volume of data to train a model to predict insurance claims fraud is important. However, they agree that there is a threshold where the size of the training data set no longer increases the accuracy (Liao and Zhu, 2014). For different machine learning algorithms, this threshold ranges from 20,000 to 60,000 transactions (Liao and Zhu, 2014).

4.1.5 Advanced Computing Requirements for Big Data

Because the definition of Big Data stipulates that there needs to be “advanced computing requirements and data-processing methods”, an evaluation of some of these methods could be beneficial. According to Zaharia et al. (2012) and Chen and Zhang (2014), cluster computing and distributed computing are a commonly used solution for large-scale analytics; cluster computing is the use of large numbers of cost-effective devices to form one single system image that can be used to solve a problem (Valentini et al., 2013). Examples of cluster computing frameworks include Hadoop (MapReduce), Storm and Drill (Dremel) as mentioned by Chandarana and Vijayalakshmi (2014), whereas Hbibbi and Barka (2016) mention MapReduce and Spark. Nunns (2016), in turn, mentions Pachyderm, Spark, BigQuery (Dremel), Presto and Hydra as cluster computing alternatives to Hadoop. From these three literature contributions, the frameworks that are commonly mentioned are MapReduce, Spark and Dremel. The current research therefore evaluates these three frameworks below.

4.1.5.1 MapReduce

MapReduce is a well-known example of an open source cluster computing programming framework that allows users to generate large sets of data, as well as process this data (Dean and Ghemawat, 2010). Advantages of using MapReduce include the fact that it is highly scalable and flexible, and it manages the partitioning of input data, scheduling of tasks, dealing with failures, as well as organising of communication between machines in the cluster (Dean and Ghemawat, 2010). It also is the most commonly used cluster computing framework as part of the Hadoop framework, which can be split up into the Hadoop File System (HDFS) and MapReduce (Gopalani and Arora, 2015). Owing to the fact it is the most widely used, support for MapReduce is higher than for

other cluster computing frameworks (Gopalani and Arora, 2015). The disadvantages of MapReduce include the fact that it is not designed for interactive workloads, but for batch workloads; also, Map and Reduce programming can be complex to achieve and the framework is slower than Spark (Gopalani and Arora, 2015).

4.1.5.2 Spark

Apache Spark is a cluster computing framework that is used for Big Data processing with a focus on streaming and near real-time processing (Jones, 2011). The advantages of Spark are that it is open source and superior to Hadoop/ MapReduce when it comes to processing speed (Shanahan and Dai, 2015). The disadvantage of Spark is that it needs the already existing HDFS file system or an alternate file system to perform analyses (Noyes, 2015).

4.1.5.3 Dremel

Dremel is a scalable implementation of a query system that was created by Google to run over trillion row tables within seconds (Melnik et al., 2010). The advantages of Dremel are that it allows users to run interactive and ad hoc queries on large sets of data within a short amount of time (Melnik et al., 2010). BigQuery is an implementation of Dremel that can be accessed via a RestFul API, without needing the hardware to support it (at a cost) (Sato, 2012). Dremel is also open source and accessed as Apache Drill (Tigani and Naidu, 2014). The disadvantages of Dremel are that it does not allow complex data-processing logic and existing data cannot be updated in Dremel (Sato, 2012).

From MapReduce, Spark and Dremel, the most apt cluster computing framework had to be chosen for this research. Although Dremel is newer than MapReduce and faster, MapReduce still has its place in the Big Data domain. According to Sato (2012) of Google, MapReduce is more fitting when performing complex data-mining operations, whereas Dremel is ideal for finding specific records with a certain set of conditions. Spark and MapReduce are consequently more fitting for this research. Although Spark is newer and faster than MapReduce, it may not be necessary to have the speed that Spark provides. The speed of Spark would only be a determining factor if the model created would need to predict fraud in near real-time – however, this research only intends to run a model weekly against the original data and then to use those results to predict fraud. Since Spark does not contain a file system, Hadoop needs to be used for the HDFS. From this comparison, it seems that Hadoop would be a good fit for the research in hand. It is therefore interesting to explore the distributions of Hadoop that could facilitate the use of Big Data.

4.1.6 Hadoop Distributions

To achieve an enterprise-quality Big Data solution that uses data science to predict insurance claims fraud, it would be recommended to use a Hadoop distribution that is used in enterprises. Examples of the most commonly used Hadoop distributions include MapR, Hortonworks and Cloudera (Gualtieri et al., 2014). To decide on which Hadoop distribution to use, an evaluation of the three most commonly used ones is performed below.

4.1.6.1 Cloudera's Open Source platform (CDH)

According to Cloudera (2017), CDH is Cloudera's open source platform that is designed for enterprise. It includes many elements that support a Big Data project. Advantages of using CDH include the fact that it has a user interface that is easily understandable (Dezyre, 2016a) and it contains many commercial add-ons for Hadoop, such as Impala and Cloudera Manager (Dezyre, 2016a; Gualtieri et al., 2014). The disadvantage, however, is that Cloudera can be seen to be slower than MapR (Dezyre, 2016a).

4.1.6.2 Hortonworks Data Platform (HDP)

According to Hortonworks (2017), the Hortonworks Data Platform (HDP) is the "only true secure" Hadoop distribution based on YARN. The advantages of HDP are that it can be used on a Windows operating system and it has a strong focus on open source additions (Dezyre, 2016a; Gualtieri et al., 2014). A disadvantage of HDP is that the user interface that is available does not have many features (Dezyre, 2016a).

4.1.6.3 MapR Converged Data Platform

According to MapR (2017), the MapR Converged Data Platform is their offering that provides Hadoop, Spark and Drill with real-time performance. Advantages of MapR are that it is seen as one of the most efficient distributions and it is more advanced when it comes to larger cluster implementations (Gualtieri and Curran, 2016). MapR also has the financial backing of Google Capital (Woods, 2014). The disadvantage is that the user interface of MapR is clearly lacking in comparison to Hortonworks and Cloudera (Dezyre, 2016a).

From the above comparison of distributions, it emerges that MapR is the best solution to be implemented in this research, due to the fact that it is the fastest implementation that is highly scalable. The other two distributions have their benefits, but the choice of Hadoop distribution can perhaps be seen as a personal choice after each one was tested.

4.1.7 Hadoop Sub-projects

Since Hadoop is a solution with many sub-projects, it was considered valuable to describe some of the sub-projects that can be used to facilitate the use of Big Data Science to predict insurance claims fraud. According to Dhyani and Barthwal (2014), the important sub-projects of Hadoop include HDFS, Hive, Pig, HBase, Cassandra, HCatalog, Lucent, Hama, Avro, Drill, Mahout, Sqoop, Flume, Chukwa, Zoo Keeper, Oozie and Ambari. In comparison, BMC (2016) mentions the four main projects of Hadoop – namely HDFS, MapReduce, Yarn and Common – as well as Spark, Hive, Pig, HBase, Oozie and Sqoop. Among the two literature contributions, the projects that overlapped included HDFS, Hive, Pig, HBase, Oozie and Sqoop.

These six projects are therefore described in more detail below.

4.1.7.1 HDFS

The Hadoop Distributed File System (HDFS) is the main file system that is part of the Hadoop project which spans multiple sets of low-cost machines. HDFS has a high fault tolerance, even though it can be deployed on commodity hardware. Its main feature is that it can accommodate large data sets and is easily portable (Borthakur et al., 2008). HDFS works through a master-slave architecture known as the NameNode and DataNodes. The NameNode manages the file system, whereas the DataNodes manage the storage on which they run. As this research's claims data would need to be stored somewhere, HDFS could be used.

4.1.7.2 Hive

Hive is used as a data management tool to more easily facilitate large sets of data in HDFS (Begoli and Horey, 2012). Hive is a data warehousing facilitator for Hadoop. It is used by many large organisations to manage and process data on HDFS. Hive translates SQL like language (HQL) into MapReduce operations on HDFS data sets (Huai et al., 2014). Since many data scientists are required to know SQL (Steynberg, 2016), Hive allows these data scientists to run queries on the data without knowing MapReduce. For the present research, it is proposed that Hive be used to aid in the management of claims data in HDFS.

4.1.7.3 HBase

HBase is also used as data management tool to more easily facilitate large sets of data in HDFS (Begoli and Horey, 2012). It is based on Google's BigTable and is a column-oriented NoSQL system (Saloustros and Magoutis, 2015).

4.1.7.4 Sqoop

Sqoop is an Apache sub-project that facilitates the import of data from a system. This is done quickly and efficiently without excess loads to external systems (Aravinth et al., 2015). To import the claims data into HDFS from a relational claims database that is external to Hadoop, using tools such as Sqoop could be beneficial.

4.1.7.5 Pig

Pig is an Apache project that performs data manipulation on data in Hadoop (Arora, 2017). Pig is an alternative to Java MapReduce programs and is intended to reduce the amount of time taken to write the mapping and reducing commands. Although this can be seen to be beneficial, the necessity of such a tool for this research was not immediately apparent.

4.1.7.6 Oozie

Oozie is a workflow management system used for Hadoop implementations (Islam et al., 2012). The advantage of using Oozie is that jobs performed by other Hadoop components such as Hive can be arranged in a specific order (Islam and Srinivasan, 2015). Once again, this is beneficial but did not directly seem to benefit the research in hand.

4.2 Data Science and Advanced Analytics

Definitions of data science are broad and varying, and as such they should be compared and evaluated to find the one best applicable to this research. Data science can be defined as “the application of quantitative and qualitative methods to solve relevant problems and predict outcomes” (Waller and Fawcett, 2013), as “the scientific study of the creation, validation and transformation of data to create meaning” (Walker, 2015) or as “the set of activities involved in transforming collected data into valuable insights, products, or solutions” (Chou et al., 2014). There is overlap among these definitions, for instance each definition has data as the source or input, some sort of data-processing techniques (i.e. quantitative and qualitative methods, validation and transformation) and there is some sort of knowledge as an output.

Based on these definitions the following definition is proposed:

4.2.1 Definition: Data Science

Data science is the science of taking data as an input, using intelligent processing methods on this data and having insight or meaning as an output.

Data science and advanced analytics are not processes on their own, but rather a collection of techniques. Research performed by Dhar (2013), Russom et al. (2011) and Loukides (2011) – all of which is highly referenced – mentions techniques used in data science. Russom et al. (2011) relate data science (analytics) to Big Data, which also adds value in this research. They also maintain that data science includes techniques such as artificial intelligence, natural language processing, data mining, predictive analytics and statistics. There is some overlap with Dhar (2013) who argues that data science includes artificial intelligence, databases, mathematics, machine learning, predictive analytics and statistics. In comparison to this, Loukides (2011) mentions data cleaning, manipulation and data quality, natural language processing, machine learning, artificial intelligence, new database and SQL techniques, statistical analysis and data visualisation as components of data science, while data cleaning, manipulation and data quality are seen as data pre-processing methods. Considering these three literature contributions, the common techniques of data science are artificial intelligence and statistical analysis. Since both Dhar (2013) and Loukides (2011) mention machine learning and complex SQL, these can be added to the list of techniques in data science. Predictive analytics is mentioned by Russom et al. (2011) and Dhar (2013), while Russom et al. (2011) and Loukides (2011) cite natural language processing as important techniques in data science and hence justify their inclusion in the list. Although only Loukides (2011) mentions data pre-processing methods, the researcher believes that these are valuable because most of these techniques cannot be used unless data has been pre-processed. Thus, the techniques of data science can include but are not limited to, artificial intelligence, natural language processing, data visualisation, complex SQL, machine learning, predictive analytics, statistical analysis and data conditioning.

Based on the above description of the techniques in data science as well as the components of Big Data, the researcher compiled the diagram in Figure 4.2 to give a detailed overview of the components of Big Data Science. This diagram was generated by listing (on the left-hand side) the components of Big Data, namely Volume, Velocity and Variety which can be incorporated into solutions that use the techniques of data science (on the right).

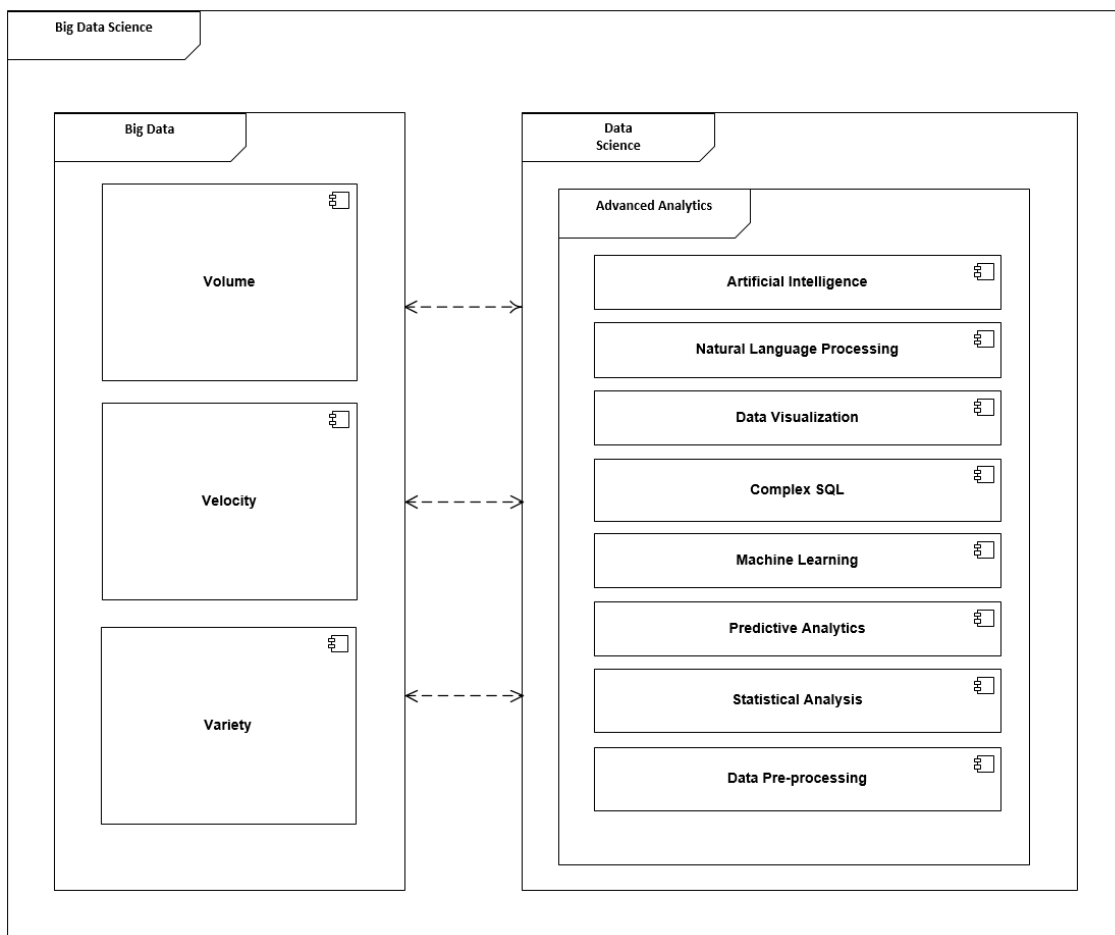


Figure 4.2: Big Data Science Component Diagram Derived from Dhar (2013), Russom et al. (2011) and Loukides (2011)

As it was previously stipulated that machine learning as a technique of data science is the intended method of predicting insurance claims fraud in this research, it was interesting to note that other techniques mentioned in Figure 4.2 overlap with machine learning. From the techniques shown in Figure 4.2, predictive analytics and artificial intelligence can be seen to have machine learning as a foundation. According to Stamford (2010), predictive analytics is one of the biggest technologies driving property and casualty insurance. Most traditional analysis techniques involve the use of historical data to analyse trends and they report on the past, whereas predictive analytics predict future trends or events. Predictive analytics can be used as a data science technique to identify claims that must be further investigated for fraud, and to reduce the number of claims that need not be investigated for fraud (Nyce and CPCU, 2007).

Since machine learning can be used to predict outcomes (Siegel, 2016), it is a building block of predictive analytics. Artificial intelligence also has machine learning as its foundation. Jordan and Mitchell (2015) maintain that machine learning is at the core of both data science and artificial intelligence. Artificial intelligence is any system that mimics human thinking and intelligence by learning from humans' thinking and intelligence (Čerka et al., 2015). Although data mining is not included in Figure 4.2,

Russom et al. (2011) identify it as a data science technique and Witten et al. (2016) stipulate that machine learning is a fundamental element of data mining. According to Fayyad et al. (1997), data mining is the use of algorithms (such as machine learning techniques) to extract patterns from data.

4.2.2 Machine Learning

Because there is consensus that machine learning lies at the core of predictive analytics, artificial intelligence as well as data mining, a better understanding of machine learning is necessary. Machine learning is the field of computer science that focuses on modelling learning in computers (Carbonell et al., 1983). According to Lloyd et al. (2013), there are two types of machine learning, namely unsupervised and supervised machine learning. These two types of machine learning are described below and comparisons are drawn between the techniques that have been used to facilitate each of them.

4.2.2.1 Supervised Machine Learning Techniques

Supervised learning models are created when there is knowledge of the input and output data of a modelled process (Nelles, 2013). The correct classification of the results is already known (Sathya and Abraham, 2013). Examples of supervised learning techniques can be C4.5, support vector machines, neural networks, AdaBoost, logistic regression or Naive Bayes, to name a few. This non-exhaustive list of examples was determined based upon the fact that the techniques were mentioned by both Orriols-Puig et al. (2008) and Caruana and Niculescu-Mizil (2006), and they were part of the top ten algorithms used by data mining (Wu et al., 2008). Artificial Neural Networks (ANNs) were initially not part of this list, but as they were important for categorical data analysis (Agresti and Kateri, 2011), they were acknowledged as an effective fraud prediction method in the previous chapter. A selection of these supervised machine learning techniques are discussed below to determine whether they would be fitting in a system that could intelligently predict insurance claims fraud. Although Chapter 3 discussed the existing literature that contained such techniques, it would be a shortcoming of this research to not consider the top techniques individually.

C4.5

C4.5, a supervised learning technique that generates a decision tree or rules algorithm from data that has already been classified (Quinlan, 2014), was developed and published by J. Ross Quinlan in 1993 (Quinlan, 1993). The main advantages of the technique are that the results are easily interpretable and it performs relatively fast for a machine

learning technique (Li, 2015). The disadvantage however is the fact that separate training sessions are needed for each attribute (Lakshminarayan et al., 1996). As a result of this, and because a claims data set would have a large number of attributes, it seems that C4.5 would not be the best fit for a system that intelligently predicts insurance claims fraud.

Support Vector Machines

Support Vector Machines (SVMs) were developed between 1962 and 1964 by Vapnik and Chervonenkis (1964). They constitute a classification and regression technique that splits a data set by creating a line known as a hyperplane between the classified parts of the data. The technique also tries to optimise the distance between this line and the sets of data (Meyer and Wien, 2015). An advantage of Support Vector Machines is that along with C4.5, this technique is often tried first, as both are easy to create (Li, 2015). Disadvantages, however, include the fact that SVMs do not scale well and that they are difficult to interpret (Wu et al., 2008). Because the research in hand focuses on Big Data in insurance, SVMs would not be the best supervised machine learning technique for this research as it would need to scale.

AdaBoost

AdaBoost was developed by Yoav Freund and Robert Schapire in 1995 (Freund and Schapire, 1995). AdaBoost is a classification technique that takes training data and generates a classification rule (Zhu et al., 2009). New data lines are then classified to existing classes when added to the technique. Advantages include the fact that the technique is relatively easy to implement, it works well with most types of data and it is easily adaptable (Zhu et al., 2009). A disadvantage, however, is that it is susceptible to noisy data (Frénay and Verleysen, 2014). If the data set that is used to train a system to intelligently predict insurance claims fraud includes high volumes of data (e.g. of many insurers and brokers), it is highly likely that the data will be noisy and hence AdaBoost would not be an ideal fit for this research. Big Data is very likely to be noisy.

K-Nearest Neighbours

The rule upon which K-Nearest Neighbours is based was developed by Evelyn Fix and Joseph Hodges Jr in 1951 (Fix and Hodges Jr, 1951). K-Nearest Neighbours (KNN) is a classification technique that takes a training set of data and then classifies new input data into groups that are unlabelled. The method of classification is lazy as the training set is used only when new unlabelled records are inputted (Peterson, 2009). The main

advantage of KNN is that it is one of the simplest machine learning techniques to implement (Imandoust and Bolandraftar, 2013). Unfortunately KNN can be slow and it is susceptible to outliers that contribute to the means (Wu et al., 2008). The latter disadvantage is the reason why KNN would not be the best fit for the current research – Big Data that comes from multiple insurers and brokers would inevitably contain many outliers that could skew the results.

Naive Bayes

According to Panda and Patra (2007), Naive Bayes is based on the Bayes Theorem that was developed by Thomas Bayes between 1702 and 1761. Naive Bayes is a classification technique where the class with the highest likelihood is assigned to a record that is shown as a feature vector (Rish, 2001). The main advantage of Naive Bayes technique is its remarkable accuracy (Al-Aidaros et al., 2010). The disadvantages however are that it is sensitive to irrelevant attributes and it assumes that predictors are independent (Bede, 2017). Since the research would like to investigate information from any attribute possible in a Big Data insurance claims data set, Naive Bayes does not fit the use case entirely. It does however have a strong advantage and as such could perhaps be useful in future research.

Classification and Regression Trees

Classification and Regression Trees (CART) were developed by Leo Breiman, Jerome Friedman, Richard Olshen and Charles Stone in 1984 (Stone et al., 1984). They are machine learning techniques that iteratively partition data (Steinberg and Colla, 2009). The result is a decision tree that can in theory determine a result. The main advantage of CART is that it can easily be interpreted by non-data scientists (Questier et al., 2005). The main disadvantage however is that it can only be used for a finite number of unordered values as it may become unstable (Timofeev, 2004). CART will not work, as the solution needs to incorporate Big Data.

Logistic Regression

Cramer (2002) maintains that logistic regression was invented by Pierre-Francois Verhulst between 1838 and 1847. Logistic regression is a predictive analytics mathematical model solution that predicts a binary outcome from a set of independent variables (Agresti and Kateri, 2011). The advantages of logistic regression include its wide availability in statistical software and the fact that it gives explicit probabilities of events occurring as an output (Westreich et al., 2010). Unfortunately, it needs a large sample size to be accurate and assumes that the data follows a log-linear shape – which may

be problematic (Westreich et al., 2010). However, these advantages and disadvantages correspond with this research, as “Fraudulent” and “Not Fraudulent” could be the binary output and it would be beneficial to have a prediction of this value. Because the current research focuses on Big Data and insurers have large data sets, the fact that the algorithm needs large sample sizes would not be a problem at all.

Artificial Neural Networks (ANN)

The first Artificial Neural Network was developed by Warren McCulloch and Walter Pitts in 1943 (McCulloch and Pitts, 1943). Artificial Neural Networks (ANN) are parallel learning systems that attempt to use organisational systems in humans to create processes and interconnections for data (Gupta, 2013). The advantages of ANNs include the fact that they are relatively easy to implement, their performance is good for continuous data, they have a lower generalisation error than decision trees, and they are able to determine interactions among variables (Tu, 1996; Dreiseitl and Ohno-Machado, 2002). However, they are often used when simpler linear regression models would suffice, they are susceptible to overfitting and they often require large computational resources compared to other techniques (Tu, 1996). Although ANNs show promise with regard to the prediction of insurance claims fraud, the large computational requirement could be problematic with Big Data.

4.2.2.2 Supervised Machine Learning Techniques: Summary

From among this list of supervised machine learning techniques, C4.5, Support Vector Machines, AdaBoost, Naive Bayes, K-Nearest Neighbours and Classification and Regression Trees have proved to be not suitable for this research as their disadvantages restrict their usage. In contrast, Artificial Neural Networks and Logistic Regression would be suitable for this research as these techniques are commonly used when predicting fraud. This fact complements our determination of possible supervised machine learning techniques that can be used to intelligently predict insurance claims fraud.

Although both neural networks and logistic regression have strong advantages and disadvantages, logistic regression applies more specifically to this research (i.e. it can give an explicit probability of fraud and requires less computational resources compared to neural networks) and as such it is given preference. Since insurance claims fraud can be stipulated as binary, since there are many independent variables and since the research focuses on Big Data, logistic regression is apt for this use case. According to Phua et al. (2010), there is a greater focus on the complex machine learning techniques such as neural networks than on logistic regression. This is misleading, as although logistic regression is less sophisticated than neural networks, it is as effective (if not more effective).

4.2.2.3 Unsupervised Machine Learning Techniques

In contrast to supervised learning, unsupervised learning involves the creation of a predictive model where the results set is not known or utilised (Nelles, 2013) and hidden patterns are found in the data. This is beneficial, as previous bias and information do not affect the results of the model (Sathya and Abraham, 2013). Examples of unsupervised learning techniques are Clustering, Expectation Maximisation and association rules. This non-exhaustive list of examples was decided upon based on the fact that they were mentioned in the top ten algorithms in data mining (Wu et al., 2008), as well as mentioned by Lingaraju et al. (2013). A selected number of these unsupervised machine learning techniques are described below.

Apriori Association Rules

Apriori association rules were developed by Rakesh Agrawal and Ramakrishnan Srikant in 1994 (Agrawal et al., 1994). The Apriori technique learns the association and correlation between fields in a data structure, and it is often used to determine what items are commonly purchased together at shops (Borgelt and Kruse, 2002). Advantages of the Apriori association rules technique include its being surrounded by a large amount of documentation, which makes it easier to understand, and there are many derivatives of the technique that can be used (Li, 2015). Its main disadvantage, on the other hand, is that the Apriori technique is memory intensive and at times also time intensive (Lin, 2014). This disadvantage can be overcome with implementations of Apriori that use MapReduce, which results in the technique having a more scalable implementation (Rong et al., 2013). As it was previously determined that fraud prediction can be performed by showing rules that are broken, Apriori association rules can be expected to work well with insurance claims fraud prediction. The disadvantage of scalability is shown to be mitigated and hence the algorithm can work with Big Data.

Expectation Maximisation

Expectation Maximisation (EM) was described and named by Arthur Dempster, Nan Laird and Donald Rubin in 1977 (Dempster et al., 1977). EM is a statistical model that determines the likelihood of something occurring and the parameters of this model itself (Moon, 1996). The advantages of EM are that it is achieved without much effort and can be used even with missing information (Li, 2015). A disadvantage however is that the expectation part of the technique can be very slow and complex (Levada et al., 2011), and hence it becomes problematic as the size of the data set increases. As the current research intends to use large claims data sets, EM would not work well for the insurance claims fraud prediction use case.

K-Means Clustering

The algorithm behind K-means clustering was developed by Stuart Lloyd (Lloyd, 1982) already in 1957, but the technique was named by James MacQueen in 1967 (MacQueen et al., 1967). K-Means is a clustering technique that takes a declared number of groups (K) and a set of data and then splits this data into the K number of groups, based on similarity (Kanungo et al., 2002). The advantages of K-means clustering are that it is easy to implement and probably more efficient than other clustering techniques such as hierarchical clustering (Islam and Ahmed, 2013). The disadvantages however are that the number of clusters needs to be specified before training and the cluster centres are biased to the initial centres (Li et al., 2015). As K-Means clustering is efficient, it should work well with Big Data. It might however be problematic to apply K-Means clustering to fraud prediction.

4.2.2.4 Unsupervised Machine Learning Techniques: Summary

From the above list of unsupervised machine learning techniques, K-Means clustering and Apriori association rules were identified as possible unsupervised machine learning techniques that could be used in an intelligent solution to predict insurance claims fraud in large sets of data. Because the disadvantages of Expectation Maximisation outweighed its advantages, EM was not suitable for this research. The existing research explored the prediction of insurance claims fraud based on supervised machine learning techniques, but it appeared that combining supervised and unsupervised machine learning techniques could increase fraud prediction accuracy. Therefore, adding K-Means clustering or Apriori association rules to a system could increase the system's accuracy in predicting insurance claims fraud.

The fact that Apriori has many implementations has great value and its disadvantage of being slow can be overcome with implementations that use MapReduce – which results in the algorithm having a more scalable implementation (Rong et al., 2013). Although K-Means clustering is ideal for outlier detection, the association rules that Apriori produces can also achieve this and can show relationships between variables. It can therefore be said that, from the algorithms evaluated, Apriori would be the best unsupervised algorithm for this research.

4.3 Data Science Platforms

According to Jovic et al. (2014) and KDnuggets (2015), Orange, R, RapidMiner, SQL, Python, Excel, Knime, Hadoop, Tableau, SAS base, Spark and Weka are the most commonly used tools (platforms) for data mining and machine learning. Wimmer and Powell (2016) in turn mention RapidMiner, Weka, Orange, R, Knime and Tanagra as

commonly used tools for data science. The overlap between the two lists are Orange, R, RapidMiner, Weka and Knime. In fact, at the start of this research, the top three data science platforms by market share were R, RapidMiner and Knime (KDnuggets, 2015). The following subsections draw a comparison between these three to determine which platform will be best suited for this research.

4.3.1 KNIME

The Konstanz Information Miner (KNIME) is an open source data science, integration and reporting platform (Berthold et al., 2009). The advantages of KNIME are that it provides a graphical interface that allows easy visualisation of the extraction, transformation and loading (ETL) process and can be integrated easily with other data science software (Rangra and Bansal, 2014). The fact that KNIME can connect to Hadoop is also beneficial for research with a focus on Big Data (Minanovic et al., 2014). The disadvantages of KNIME are that it has limited methods for error measurement and is not as widely used as R and RapidMiner (Rangra and Bansal, 2014; KDnuggets, 2015).

4.3.2 R

R is statistical software that can be used for a variety of data science and machine learning tasks (R Core Team, 2000). The advantages of R include its wide usage and hence its adequate support. It also has a strong link to academia and as such, has a large focus on machine learning. It furthermore contains very fast implementations of data mining algorithms (Jovic et al., 2014), and according to Dezyre (2016b), R and Hadoop are a great combination for Big Data. A main disadvantage of R is that it may be a problem to learn how to use the array-based language (Rangra and Bansal, 2014).

4.3.3 RapidMiner

RapidMiner is software used for machine learning, predictive analytics and data mining (Zheng and Dagnino, 2014). The main advantage of RapidMiner is that it contains many built-in procedures for integration, transformation and data analysis (Rangra and Bansal, 2014). RapidMiner can also be used for integration with Hadoop (Prekopcsak et al., 2011). The disadvantage of RapidMiner is that it is too focused on SQL manipulation (Rangra and Bansal, 2014).

From the comparison, R can be seen to add great value to data science. Since R can be coupled with Hadoop and Spark (sparklyr, 2016; Gollapudi, 2016), it is an attractive option when utilising data science with Big Data. With the additional fact

that R is used by academics and vast research is conducted into the use of Data Mining algorithms in R, R is considered the best option for this research.

4.4 Data Science: Privacy Consideration

The research contained in this dissertation is specifically focused on the fraud committed in property and casualty insurance claims and predicting this fraud while considering privacy legislation. Thus, the use of data science with Big Data to predict insurance claims fraud could be improved upon by incorporating privacy considerations.

An approach that needs to be applied in the prediction of insurance claims fraud through Big Data and data science is one that focuses on protecting the privacy of policyholders whilst performing the machine learning techniques and storing the data in a Big Data structure. This approach to data mining that protects privacy is known as Privacy Preserving Data-Mining (PPDM) (Xu et al., 2016). Although the research reported on in this dissertation does not restrict the data science methods used to mine data, PPDM can be seen to be applied to other data science techniques. The advantages of using PPDM include the protection of the people's privacy, the simplicity of its implementation based on the method used and the fact that there are different methods of applying PPDM (Laskar et al., 2014). The disadvantages of using PPDM however, may include information loss and the fact that privacy protection is not guaranteed (Laskar et al., 2014).

Examples of PPDM methods can include (but are not limited to) anonymisation, perturbation and cryptographic methods (Rajesh et al., 2016; Taric and Poovammal, 2017). Perturbation involves adding "noise" to a data set by modifying the data set through the addition of randomness to individual values (Taric and Poovammal, 2017). An example of this would be to add Gaussian noise to personal information such as telephone numbers (Kalaivani and Chidambaram, 2014). Policyholders would no longer be identifiable by their telephone number if Gaussian noise were added to it. Anonymisation involves generalising and suppressing records in a data set so that individuals can no longer be identified, but the data can still be used for data-mining methods (Parmar and Shah, 2016). An example of this would be to condense personal information into a range of values (Jaiswal et al., 2016). Instead of showing a policyholder's telephone number in the data set, it can display the first three digits and replace the remaining characters with '*'. Cryptographic techniques involve the encryption or cryptographic hashing of personal data so that the data can be mined across distributed parties (Patilwar and Agrawal, 2017; Bansal et al., 2017). This could for example involve hashing all policyholder telephone numbers with the MD5 algorithm. Since there are multiple insurers and brokers, and since cryptographic PPDM is best for collaborative data mining (Rathna and Karthikeyan, 2015), a cryptographic PPDM technique could add

value to an intelligent way of predicting insurance claims fraud within the restrictions imposed by privacy legislation. An example of this would be the encryption or cryptographic hashing of any information that can uniquely identify a person (Chakravorty et al., 2013). If data no longer has value for an attacker because it has been hashed enough to conceal policyholders' information, then the associated cyber security risk can be considered somewhat mitigated.

This research will incorporate the use of PPDM into the application of Big Data and data science to predict insurance claims fraud. Although there is evidence to substantiate the use of PPDM (Bertino et al., 2005; Liu et al., 2006; Xu et al., 2014), there is a lack of research applying PPDM directly to the property and casualty / short-term insurance domain.

4.5 Discussion

This chapter described Big Data Science by formulating definitions of Big Data and data science and showing how these two terms are interdependent. It also described the three V's of Big Data, namely volume, velocity and variety, and discussed Big Data Frameworks, the techniques of data science and data science platforms. The chapter described the difference between supervised and unsupervised machine learning and explained techniques used for each of these machine learning types. The following key points can be noted from the chapter:

- Logistic Regression can be seen as an ideal supervised machine learning technique to predict insurance claims fraud. It was determined in Chapter 3 that Logistic Regression is commonly used to predict fraud in financial systems and this chapter reinforced this notion as it was shown as an appropriate supervised machine learning technique for predicting insurance claims fraud.
- Apriori association rules is an ideal unsupervised machine learning technique to predict insurance claims fraud, since it can be adjusted to work with large sets of data and it derives rules.
- Using R in combination with the MapR distribution of Hadoop constitutes ideal tools to use for this research.
- The appropriate size of the data set for machine learning ranges between 20,000 and 60,000 transactions.
- Privacy Preserving Data Mining can be used to supplement data science techniques to protect the privacy of insurance policyholders. Cryptographic hashing will be useful if data is to be shared between insurers and brokers.

A broad understanding of these fields should result in a better understanding of how they will be applied to an intelligent system that can be used to predict insurance

claims fraud.

From this chapter and the previous chapters, background has been presented as to the necessary requirements and components of a system that can intelligently predict insurance claims fraud. This needs to be further specified and refined so that it can be understood *what* precisely needs to be performed to intelligently predict insurance claims fraud and *how* this can be achieved. To do this, the research in hand puts forth a framework, architecture and model. As will be shown, a framework defines the interfaces, rules and services that are required to perform a set of tasks (Kopper, 2009). In comparison, an architecture specifies the design, logical organisation as well as conceptual structure of a system (Oxford English Dictionary Online, 2017b). Lastly, a model is the theoretical or empirical description of a system regularly expressed using mathematical notation so as to understand prediction, calculations etc. (Oxford English Dictionary Online, 2018).

This framework, model and architecture are shown in Figure 4.3. Figure 4.3 shows how the three components of this solution fit together. The differences in these components and the way in which they relate to each other were derived from research by Nilsen (2015) and Morgenstern et al. (2017). The framework shows *what* needs to be performed to intelligently predict insurance claims fraud and the model and architecture show *how* this can be achieved. The model and architecture are derived from this framework as it is first necessary to understand what needs to be done before it can be shown how to achieve the required result.

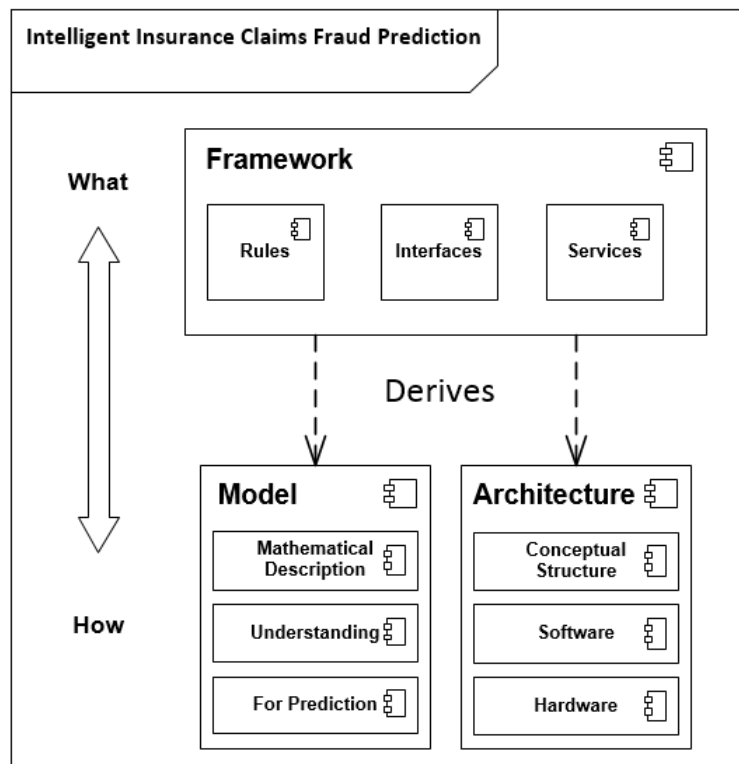


Figure 4.3: The Framework, Model and Architecture of this Research

5 Proposed Framework for Intelligent Insurance Claims Fraud Prediction

Chapter 4 described Big Data and data science as terms that are closely connected. A description of each term was given to gain a theoretical understanding of the fields of Big Data and data science, and to show that they can together provide a solution to intelligently predict insurance claims fraud.

It is the intention of Chapter 5 to show how such a purpose can be achieved by proposing the components of a framework that can be used to intelligently predict insurance claims fraud.

The chapter begins by describing the high-level requirements of the proposed system and follows on to develop a suitable framework that includes services, rules and interfaces. This is followed by a use case diagram that shows the application of this framework and how it fits into the insurance industry.

5.1 Requirements

A number of requirements have been identified based on information gathered in the previous chapters. These requirements must be met to derive a framework for intelligent insurance claims fraud prediction. The most widely used standard for describing requirements is the Requirements Classification Schema described by the Business Analysis Body of Knowledge (BABOK) (Brennan et al., 2009). Since it is the globally accepted standard, it was used to structure the requirements discussed in this section (Masters, 2009). The types of requirements described are business requirements, stakeholder requirements, transition requirements, and solution requirements. Solution requirements can be further divided into functional and non-functional requirements. According to Brennan et al. (2009), these five types of requirements can be described as follows:

- Business requirements are the goals and objectives of the system.
- Stakeholder requirements are the requirements of the stakeholders that need to be met to fulfil the business requirements.

- Transition requirements are the needs that must be met in order for the system to move from the current state to the future state (they will no longer exist once the system has been implemented).
- Functional requirements describe what the system should be able to do.
- Non-functional requirements describe the conditions that must exist for the system to achieve the functional requirements.

The high-level requirements for this framework that are described below do not include all possible requirements; they rather serve to give an overview of its most important requirements.

5.1.1 Business Requirements

- **REQ 1:** A framework must be designed to be used to intelligently predict insurance claims fraud.

5.1.2 Stakeholder Requirements

- **REQ 2:** The framework must predict fraud after an insurance claim has been assessed, but before it is investigated. Once the claims handler (as one of the stakeholders) has processed a claim, the framework must predict whether the claim is fraudulent or not, and if necessary, it can then be investigated by a claims investigator (another stakeholder).

5.1.3 Transition Requirements

- **REQ 3:** Historic claims data must initially be added as a bulk upload to train the system.
- **REQ 4:** Existing processes of insurance claims investigation must be improved when moving from the current to the new state of insurance claims processing.

5.1.4 Functional Requirements

- **REQ 5:** The framework must use intelligent methods such as machine learning to intelligently predict insurance claims fraud. The machine learning will use both supervised and unsupervised machine learning techniques, namely Apriori association rules and logistic regression. These two techniques were determined to be ideal for the prediction of insurance claims fraud in the literature study of this research in Chapter 4.

- **REQ 6:** The framework that is used to intelligently predict fraud must have explanation ability so that it can be used by forensic accountants and insurance claims investigators.
- **REQ 7:** The framework must protect the privacy of policyholders.
- **REQ 8:** Because insurance claims data is acquired from a variety of sources, all data must be cleaned before it is added to the Big Data file system.

5.1.5 Non-functional Requirements

- **REQ 9:** Because the insurance industry deals with large data sets, the framework must accommodate the three V's of Big Data – it must cope with a high volume of claims records, the velocity of claims processing must be fast and there must be the option of using a variety of data types and sources for the claims.

5.2 Framework

From the requirements mentioned, a framework can be derived. According to Kopper (2009), a software framework defines the interfaces, rules and services that are required for it to perform a set of tasks; it is a general or high-level design. In the case of a qualitative analysis, a conceptual framework is a graphical representation of variables, constructs and factors and the relationships between them (Miles and Huberman, 1994). According to the Longman Dictionary of Contemporary English (1995), a framework contains the rules, beliefs and ideas whereby something is created or by which decisions can be made. Phan et al. (2001) maintain that frameworks show the relationships between objects.

These four contributions of what constitutes a framework are similar, but they do not provide any exact parallels. It is clear, however, that frameworks contain objects or components and that there are relationships or interfaces between these components. Kopper's (2009) view on what constitutes a software framework is closest related to this research and as such was used to derive what is required. The diagram in Figure 5.1 is proposed on the basis of the research that was conducted so far. The diagram is a high-level representation of a framework and what a framework should include. It does not show the low-level requirement such as rules, as this would over-complicate the diagram, but instead includes the components/objects, services/interfaces and how these relate to one another.

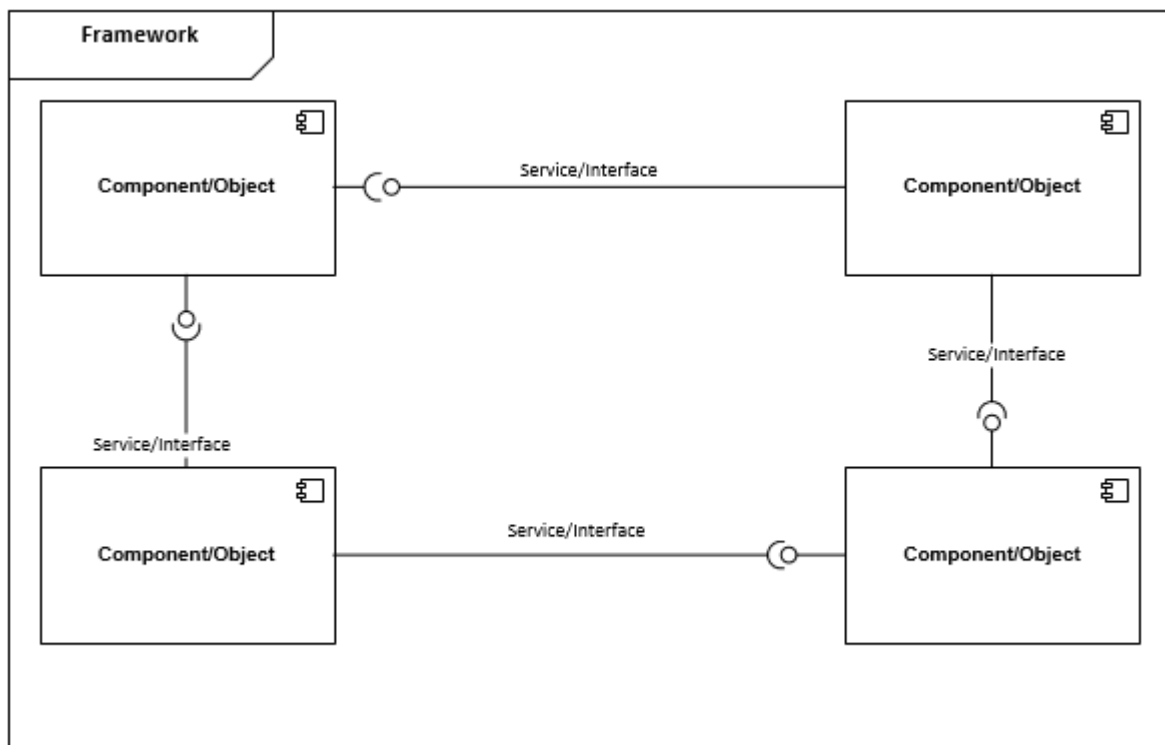


Figure 5.1: The Necessary Components of a Software Framework from Research by Kopper (2009) and Phan et al. (2001)

In this research, the task was to intelligently predict insurance claims fraud by proposing a suitable framework. Therefore, the interfaces, rules and services of this framework need to be explained. The interfaces can be derived from the required services and hence the services are described first.

5.2.1 Services

A service is a group of functionalities provided by a system and it can also contain other services (Sturm and Shehory, 2004). Since the purpose of this research was not to develop an end-to-end system but rather to build a sub-section of this system, a limited number of services were required. The services for this framework included the following:

Service 1: Clean the data for machine learning methods.

Service 2: Protect the privacy of policyholders through PPDM.

Service 3: Predict insurance claims fraud through unsupervised machine learning.

Service 4: Predict insurance claims fraud through supervised machine learning.

Service 5: Provide explanation ability to be used by forensic accountants.

5.2.2 Interfaces

Interfaces are descriptions of public methods that define interactions between a request and a service. The methods in the intelligent fraud predictor interface derived from the services above are defined here, but they were not implemented for the interface (Brown et al., 2002). For the purposes of this research, the methods in a framework that could intelligently predict insurance claims fraud would include:

```
1 public interface IIntelligentFraudPredictor
2 {
3     List<Claim> getClaimsDataFromBigDataStore();
4     List<Claim> getCleanedData();
5     List<Claim> getEncryptedData();
6     void trainIntelligentFraudPredictorModel();
7     decimal getSupervisedFraudPrediction();
8     decimal getUnsupervisedFraudPrediction();
9     decimal getCombinedFraudPrediction();
10    string getFraudPredictionExplanation();
11 }
```

5.2.3 Rules

The rules of the framework constrain the design of the system and they impose a structure on the system (Gaweł and Skalna, 2014). Therefore, the rules would not necessarily be shown in a diagram that depicts the framework. For the framework that is proposed, the rules can be described as follows:

Rule 1: The size of the data set to train the machine learning model must be between 20,000 and 60,000 transactions, based on the type of machine learning algorithm used, to increase accuracy.

Rule 2: Because the solution must cater for the fact that insurance companies have large sets of data, the training data set will need to come from a Big Data file system.

Rule 3: Both unsupervised and supervised machine learning algorithms must be used to intelligently predict insurance claims fraud. Using only supervised or unsupervised machine learning techniques would reduce the accuracy of the research.

Rule 4: The prediction of insurance claims fraud must not be restricted by the data structure of the claims records.

Rule 5: The solution needs to be re-trained regularly to maintain accuracy.

These rules, services and interface, as well as the information in the previous chapters were used to compile the component diagram of the proposed framework (see Figure 5.2).

5.2.4 High-level Component Diagram Showing the Fraudulent Claims Prediction Process

A component diagram is a useful way to show the relationships between parts of a system and hence provides a high-level description of this framework (Bell, 2004). For the purposes of this research, a diagram was created with the chosen machine learning algorithms (Apriori association rules and logistic regression) as components.

Figure 5.2 covers the scope of this research and indicates the components required to intelligently predict insurance claims fraud. The left-hand side of the diagram shows the scope and main purpose of this research, while the right-hand side shows the components required to achieve this.

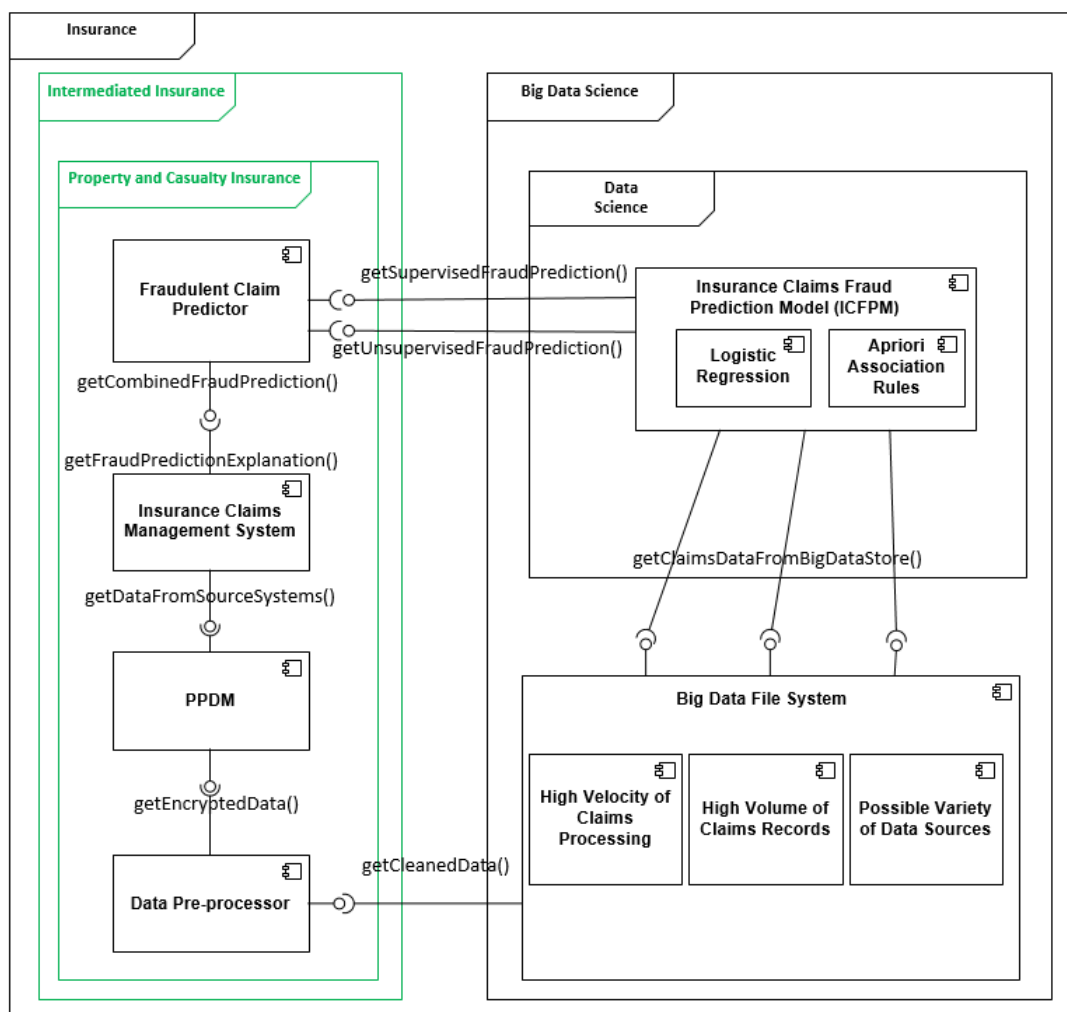


Figure 5.2: Component Diagram of the Proposed Framework for Intelligent Insurance Claims Fraud Prediction

5.3 High-level Use Case Diagram Showing the Fraudulent Claims Prediction Process

The components that should be used in an intelligent method of predicting insurance claims fraud have already been determined, but a broader understanding of where these components are positioned can be gained through a use case diagram. The design in Figure 5.3 outlines the basic use cases of a claims process that would incorporate the previously described framework to predict insurance claims fraud. Although this chapter has described in detail what is necessary to intelligently predict insurance claims fraud, a holistic view of the process could also be beneficial. A use case diagram is shown in Figure 5.3 to provide this holistic view.

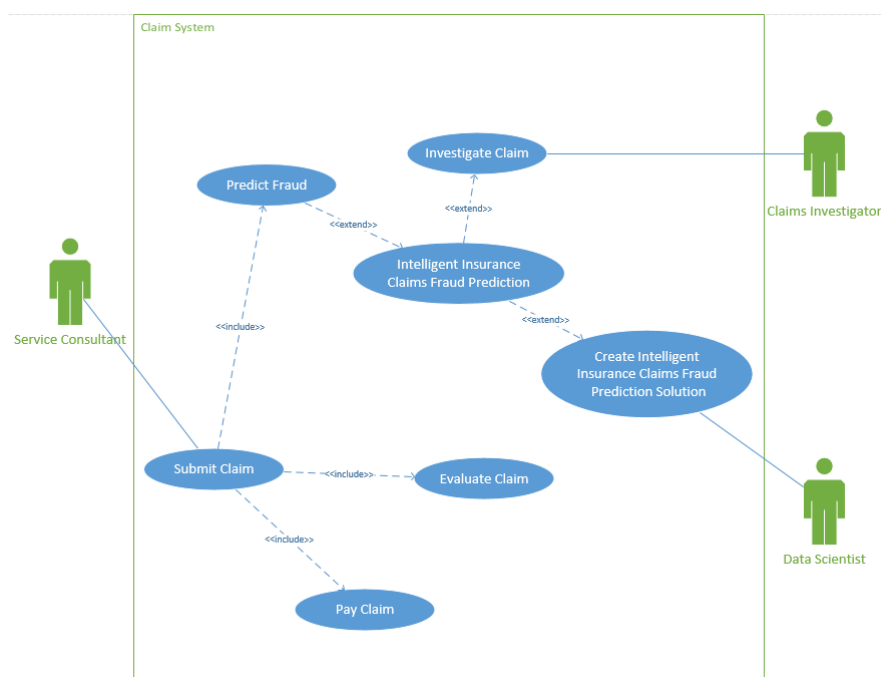


Figure 5.3: High-Level Use Case Diagram Showing the Claims Prediction Process

To understand this diagram, the actors and use cases in the diagram should be explained in greater detail. This is achieved below by briefly describing each of the actors and use cases that are necessary in a solution that intelligently predicts insurance claims fraud.

5.3.1 Actors

In the process illustrated in Figure 5.3, three basic actors are noted, namely a service consultant, a claims investigator and a data scientist.

A service consultant is processing the claim because having dedicated loss adjusters or highly experienced claims teams can prove to be expensive (ILASANews, 2014). Often,

large claims teams process all claims and only if a claim has a high value or proves to be suspicious, a loss adjuster and a claims investigator would be involved.

A data scientist would also be involved in this process. The data scientist is the person who examines the data and information and who creates a solution that can be used to make decisions (Davenport and Patil, 2012). Because an intelligent solution will involve analysing the data and using it to make a decision about whether a claim is fraudulent or not, the data scientist is vital in creating such a solution. One or more data scientists would be involved in creating and maintaining a model that uses Big Data frameworks and data science platforms.

5.3.2 Use Cases

The high-level use cases involved in this process include submitting claims, evaluating claims, paying claims, predicting fraud, investigating claims, and using and creating an intelligent solution to predict this fraud.

A data scientist would create a fraud prediction model containing sub-models that include Big Data frameworks and data science platforms which are used to predict fraud. This model would need to be re-run on a regular occasion to keep the model relevant and abreast with current trends.

The sub-models would be used to predict fraud on claims that have been submitted and if fraud is found, a claim will not be paid out.

5.4 Discussion

This chapter described a proposed high-level framework that can be used to intelligently predict insurance claims fraud. A high-level design was consequently put forward to show the basic requirements of an intelligent method of predicting such fraud. This high-level design was found to be not sufficient and hence needed to be enriched and developed to provide a better understanding of how to intelligently predict insurance claims fraud. For this purpose, a detailed design will be put forward in Chapter 6 to address the key requirements of the proposed framework:

- Insurance claims fraud must be predicted intelligently by using logistic regression and Apriori association rules.
- Since insurance claims fraud has to be explained by forensic auditors, the solution that intelligently predicts fraud must have explanatory ability.
- The design needs to cater for the fact that Big Data is used in the insurance industry and as such, the model should not be restricted by the size of the source data set.

- Since there are legislative constraints on the use of data, privacy needs to be considered through PPDM.
- Pre-processing of data must be considered to ensure that data is clean and correct.

These requirements will be achieved through a detailed design and a high-level architecture – as described in the next two chapters respectively.

6 Detailed Design

Chapter 5 described a proposed framework for intelligent insurance claims fraud prediction and stated the requirements that need to be met. The high-level design had to be further enriched to become a detailed design and achieve a better understanding of how to intelligently predict insurance claims fraud. It is the purpose of Chapter 6 to describe such a detailed design.

To gain an understanding of how to create a solution that can intelligently predict insurance claims fraud, existing processes should be examined. Eckerson (2007) maintains that to implement a sustainable predictive analytics solution, six steps need to be followed, namely project definition, data exploration, data preparation, predictive model creation, deployment and model maintenance/ management. This process complements Kurgan and Musilek's (2006) contributions on existing Knowledge Discovery and Data Mining Models research, which compare the steps involved in Data Mining and Knowledge Discovery models. Kurgan and Musilek (2006) grouped the steps under the headings of domain understanding, data understanding, data preparation, data mining, evaluation of results and deployment of results.

These literature contributions showed clear overlap, and hence the structure of this chapter could be derived. Project definition and domain understanding were excluded as these steps had previously been determined in this research. Data understanding and exploration were also dealt with in previous chapters, but some level of understanding would need to be performed in data preparation. As data preparation would be the first step in this detailed design, it makes up the first section of this chapter. Data mining and predictive model creation was merged into one step and as such, this research describes this step as model creation. Evaluation of results – which was not included by Eckerson (2007) – was added to the deployment step, which can be converted into a step of knowledge application, as the model would eventually be deployed and the results would be applied to the system. Lastly, although Kurgan and Musilek (2006) do not mention model maintenance as a step, it was previously determined in this research that regular training is an important requirement for a data science model. Therefore, the last step would be model maintenance.

The four steps described above established the structure of this chapter, as shown in the diagram in Figure 6.1. The diagram is generic as it does not apply directly to this research's problem domain. Figure 6.1 shows the sections of this chapter with

their corresponding section numbers on the right. However, the sections do not apply directly to intelligently predicting insurance claims fraud. Chapter 6 applies these sections to the requirements of the research and towards the end of the chapter, a proposed sequence diagram is portrayed which is not generic and which applies directly to this research’s use case. The example described in a previous chapter of an instance where an agent from an insurance brokerage submitted fraudulent claims without the policyholder’s knowledge is used to aid in the description of each section of this chapter (SAICB, 2014). The agent submitted claims with different account details on behalf of policyholders who had in fact not claimed from their insurers.

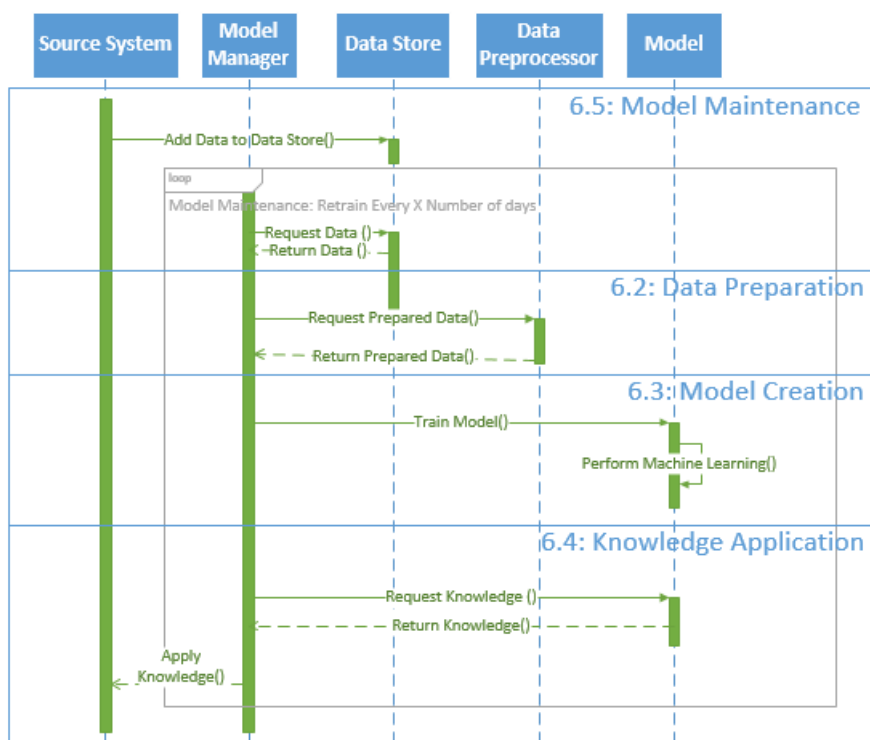


Figure 6.1: Sequence Diagram Showing the Structure of this Chapter

6.1 Data Preparation

Because this research has an element of privacy and it is intended for PPDM to be utilised, the data preparation step of predictive analytics needs to be adjusted to fit with PPDM. It was previously stated that the data preparation step includes both data exploration and preparation in this research. Data preparation also commonly includes pre-processing and transformation (Ordonez, 2011). According to Punitha and Amsaveni (2011), when PPDM is utilised, data needs to be prepared so that personal information is not uncovered during the data-mining process. It is for these reasons that the data preparation step can be further separated into seven sub-steps, namely data exploration, data pre-processing, data cryptography, data extraction, data

cleaning, data import and data transformation. These sub-steps are described below.

6.1.1 Data Exploration

Data exploration involves understanding the sources of data. These sources need to include the correct variables, history and an adequate number of records (Eckerson, 2007). The sources of data can be internal as well as external, and they need to be relevant to the task at hand (Kurgan and Musilek, 2006). Quality issues regarding the data also need to be identified during the data exploration step. This step need not be performed every time the model is retrained, unless there are valid reasons to do so, such as new fields being added, etc. Predictive analytics models are trained by narrowing down the fields that are necessary to gain an accurate result (Eckerson, 2007). The different claims fields that were described in previous chapters can be refined and filtered, depending on the use case. An example of such filtering would be to remove unnecessary policyholder information, such as race or religion.

6.1.2 Data Pre-processing

For the data pre-processing step, methods such as outlier removal, removal or repairing of null values, and fixing of discrepancies in certain fields (e.g. the policyholder's surname) are performed (Singhal and Jena, 2013). In this research, it is proposed that the processes involved in outlier removal and the removal of missing values should occur in the data-cleaning step of this framework. Discrepancies in fields such as surname must be repaired before the data is exported from the source system, as the data is intended to be hashed for PPDM. Failure to do so can result in the data being incorrect. In standard processes, pre-processing of the data by means of data cleaning should preferably occur after extraction (El-Sappagh et al., 2011). Owing to the sensitivity of insurer data and the fact that insurer data contains personal information, pre-processing is vital before the data cryptographic hashing step.

To facilitate the cleaning of data, open source tools such as OpenRefine (Ham, 2013) have been developed to make the cleaning of data less problematic and tedious. An example of a feature in a tool such as this one that can add value to this research is the repairing of field discrepancies by standardising names, insurers, brokers and claim service providers that have been inputted incorrectly or spelt incorrectly. If the names have not been standardised, the machine learning techniques will not see these values as the same and may well interpret the results incorrectly. Therefore, using such a tool or passing the fields through algorithms such as Levenshtein, Guth or NYSIIS (Snae, 2007) can result in standardised data and representation of fields.

6.1.2.1 Data Hashing

Data hashing follows after the data has been standardised. This step is important where anonymity is a key factor. As one of the core research questions in this study are whether an intelligent solution can be used to predict fraud despite the constraint of privacy, the step that performs data cryptography is highly important to the proposed solution.

The data used to train the data science model needs to be anonymised enough to ensure that policyholders cannot be identified easily. This must be done in such a manner that the data can still be utilised by machine learning algorithms. Any data that can be aggregated together to uniquely identify a person, such as postal code, date of birth and gender, needs to be anonymised (Sweeney, 2002). The research in hand suggests hashing all fields that can be used to identify a person. Hashing was chosen as the method of cryptography for PPDM because hashing is efficient, it is one-directional, it cannot be unhashed and it produces a short-lengthed digital output (Henzen et al., 2009). Although this is true with regard to hashing, other cryptographic methods could also be used.

6.1.2.2 Data Extraction

The data extraction step occurs after the data has been encrypted and pre-processed. The data is sent from the source system or database to the target data location. In the case of this research, it would involve moving the data from the claims management system to the Big Data file system. Many factors need to be considered during this step, such as the structure of the data that is being extracted, communication protocols, operating systems and software platforms (El-Sappagh et al., 2011). The extraction process should not be under-estimated as a vital component of the data science process, as many legacy systems are at play in the financial services industry. Financial service providers (FSPs) such as insurers spend substantial amounts of money maintaining these systems (Crotty and Horrocks, 2016). It is therefore important to ensure that the data extracted from these systems are in a usable and correct format.

Since the current research domain is not country specific, it is suggested that data extraction be sufficiently standard to apply to the property and casualty insurance domain in both developed and developing countries. Common standards such as the data standards described by ACORD (the Association for Cooperative Operations Research and Development)(ACORD, 2015) can be used to achieve this. Using such a data standard means that the inputted data need not be re-engineered when applying the machine learning technique.

6.1.2.3 Data Cleaning

Data discrepancies are fixed in the pre-processing phase, but because missing values are not repaired and outliers are not removed, a further data-cleaning step needs to be performed after data extraction in order to reduce the load on the claims management or source system. The original (and previously mentioned) pre-processing step is used to prevent erroneous data after hashing, but the data-cleaning step is used to perform the bulk of data fixing, removal of outliers and repair of missing information. Examples of this would be a claim with a zero-valued claim amount, or if no policyholder information is linked to a claim.

6.1.2.4 Data Import

Subsequent to the data being fully encrypted, extracted and cleaned, it can be stored. For solutions using a Big Data framework, this is done within a cluster computing file system (Zaharia et al., 2012). As mentioned earlier, cluster computing involves the use of large numbers of cost-effective devices to solve a problem (Pavlo et al., 2009). Two common examples of cluster computing models are Spark and MapReduce (Zaharia et al., 2012). Within these cluster computing models, the data is stored in a distributed file system. In the insurance industry, NoSQL and Hadoop are often used for this type of storage (Oracle, 2016). More details about such storage are provided in the architecture chapter of this research.

6.1.2.5 Data Transformation

In many extraction, transformation and loading (ETL) processes, the data transformation phase includes data cleaning (Vassiliadis et al., 2002). For this research, however, the transformation step included creating a finite set of output values from a finite set of input values and filtering the values accordingly. The research intended to prove that machine learning techniques can be used to predict fraud and, as such, the transformation that was applied would need to be effective and produce the correct finite output values.

Often-declared data transformation rules developed by machine learning techniques require that training variables must be discrete instead of continuous values (Bay, 2000). To meet this requirement, the training variables must be discretised. Thus, one would ideally change a continuous variable – such as the total policy revenue, which is a floating point number – into a factor that expresses a range.

Another requirement of machine learning techniques is that certain algorithms can cater only for continuous variables. Discrete variables such as the policyholder's name and surname would therefore be filtered out.

A final example of the transformation of data that occurs in the short-term insurance/property and casualty use case would be the creation of calculated fields. For this application, the numeric difference between the start date of a policy and the date of claim is calculated, as it is considered a valuable field.

Once the data has been fully pre-processed, the model creation can take place.

6.2 Model Creation

Contrary to Kurgan and Musilek (2006), Eckerson (2007) states that predictive model creation is an iterative process and not performed in a single phase. The process of creating one of these models involves applying a data-mining algorithm to a data set to achieve an accurate prediction of the dependent variable. This process is replicated using many combinations of variables to see which ones have the best result, and to best accomplish this, a data set is split in half – into a training data set and a validation data set. The first half is used in conjunction with the results to “train” the model, while the remaining half is used without the results to test how accurately the model predicts the dependent variable. For model creation, machine learning data science techniques are not frequently published, as they should be kept confidential for security reasons (Kou et al., 2004).

It was determined in Chapter 4 that both supervised and unsupervised machine learning would be used to train the proposed model as the two techniques would be more accurate than using only one or the other. It was also determined that Apriori association rules and logistic regression would be the best fit for this research. Logistic regression would be used as the supervised machine learning method and a fraudulent claim indicator would be the dependent variable. Apriori association rules would be the unsupervised machine learning method that would generate rules about claims. Since the training data set would contain claims that are flagged as fraudulent, rules would appear that have “FraudulentClaimIndicator=1” as the consequent (eg. “TotalPremiumRecieved=0 => FraudulentClaimIndicator=1”). Rules that do not contain the “FraudulentClaimIndicator=1” as the consequent of the rule would also appear such as “City=Pretoria => Province=Gauteng”. The rules could therefore be split into two sets, rules that have “FraudulentClaimIndicator=1” as the consequent of the rule can be classified as *supervised association rules* and rules that show normal claims patterns without any mention of fraud as *unsupervised association rules*. Model training with logistic regression and Apriori association rules is described in more detail below.

6.2.1 Model Training

To train the model, the claims data is imported from the Big Data file system, and half of a sample of claims data is used. Once the training had been completed, the second half of the sample claims data could be used to test the model. In the current research, the researcher trained the model by generating association rules and then fitting a logistic regression model to the data. This process is expanded on below.

6.2.1.1 Generate Association Rules

To generate the association rules using the Apriori algorithm, the measure of validity of a rule is regularly expressed with support, confidence and lift (McNicholas et al., 2008). These can be expressed in formulas as shown by Zhao and Bhowmick (2015) based on the original algorithm from Agrawal et al. (1994). The support of the rule is

$$support(A \Rightarrow B) = P(A \cup B)$$

which is all items with the same output and input factors that develop a rule. The support can be shown as a percentage. An example of this in the insurance domain would be the number of times claims with certain attributes happen in the full data-set.

The confidence is

$$\begin{aligned} confidence(A \Rightarrow B) &= P(B|A) \\ &= \frac{P(A \cup B)}{P(A)} \end{aligned}$$

which shows a measure of how often the rule is true. An example of this in the insurance domain would be there is a 50% confidence that the claimant is male if there is an equal share of male to female claimants.

Lastly, the lift of the rule is

$$\begin{aligned} lift(A \Rightarrow B) &= \frac{confidence(A \Rightarrow B)}{P(B)} \\ &= \frac{P(A \cup B)}{P(A)P(B)} \end{aligned}$$

which indicates the measure to which event's A and B are not independent. An example of this would be the measure to which claiming a certain value and the claim being fraudulent are independent.

For this research, lift is used as the indicator of the importance of a rule.

6.2.1.2 Fit Logistic Regression to Data

According to Kleinbaum and Klein (2010), the logistic model is

$$P(X) = \frac{1}{1 + e^{-(\alpha + \sum \beta_i X_i)}}$$

based on the original research of Pierre-Francois Verhulst which in this case can be seen as the conditional probability of a claim being fraudulent. The α and β_i are unknown values that need to be estimated from a list of claims that will train the model.

It needs to be specified that to train the logistic regression model, there are assumptions that must be met. The following assumptions must and are met (Sibanda and Pretorius, 2012):

1. The outcome must be dichotomous or discrete which is met owing to the fact that a claim is either fraudulent (1) or not (0).
2. There should be no outliers in the data-set which can be met by converting values to Z-scores and items that have a Z-score of greater than 3.29 or less than -3.29 are removed (Jauhar et al., 2016).
3. There needs to be a sufficient number of responses to reduce standard errors which is the case as this research has a focus on Big Data.
4. The multicollinearity must not be high which can be seen by the fact that inter-correlations amongst predictors are not high.

6.2.2 Accuracy Testing

After the model has been created, it can be tested. This process involves removing variables that tend to skew results or are unnecessary. Claims that exist with prior knowledge that they are fraudulent can be appended to the model to test whether they have correctly been predicted as fraud. This gives valuable insight into whether the model is effective and what needs to be altered to make predictions more accurate. Commonly used tests for statistical accuracy in machine learning include Receiver Operating Characteristics (ROC) analysis, Precision, Recall and F-measure (Powers, 2011). Although it has been mentioned that ROC analysis, Precision, Recall and F-measure can be biased, determining the correctness of an accuracy test is considered as beyond the scope of this research and as such, once the model has been trained, the most fitting accuracy test can be applied.

6.3 Knowledge Application

Once knowledge has been gained from a model, further systems can be created to gain competitive advantage. Analysts will show the knowledge and rules generated from these systems in the forms of ratings, concept maps and rules (Schiuma et al., 2012). d'Aquin and Jay (2013) suggests that the goal of knowledge elicitation is to supply end-users of systems with automatic recommendations.

The application of knowledge (after it has been extracted) to new and existing systems is a vital step in a solution that can intelligently predict insurance claims fraud. Instead of analysts creating rules and concept maps from the knowledge gained from the model, the automatic creation of rules that can be applied can be extremely useful.

6.3.1 Creation of XML Rules based on Apriori Association Rules

Often, the rules that are generated through models are not usable by end-user systems. It is therefore proposed that such rules be transformed into XML, where they can be imported into the system and applied to current claims. This will have the benefit of systems not needing to be rebuilt due to hard-coded business rules. An example of this proposition would be having an XML rule that shows that any claim that has been submitted by the broker concerned and that has the specific bank account details of this broker, must be fraudulent, or any claim paid into that account must be fraudulent.

Once these rules have been imported into a claims management system, automatic recommendations can be made.

6.3.2 Addition of Logistic Regression Model to Systems

After a logistic regression model has been fitted to the data, the logistic regression model needs to be added to systems so that the latter can run new claims through the model and predict an outcome (i.e. whether the claims are fraudulent or not).

6.3.3 Automatic Recommendations

Once the association rules have been mapped to XML and imported into a system, recommendations can be made based on new claims. If the association rules have a fraudulent claim indicator as their antecedent (result), they can be used to indicate whether a claim is fraudulent. If a rule is broken, this is indicated too. The prediction of a fraudulent claim through logistic regression is also added to this process (see Figure 6.2).

Variables are described in Table 6.1 to provide a better understanding of this process. The table shows a key of acronyms used in the sequence diagram with a variable name and a description of the variable.

Table 6.1: Description of Variables Necessary to Predict Fraud Using this Research's Approach

Key	Variable Name	Description
RBVP	Rule Break Value Percentage	For every unsupervised machine learning rule that is broken (rule with FraudulentClaimIndicator not in the result), the rule break value percentage is added to the total score showing whether a claim is fraudulent. The RBVP is a benchmark percentage and decreases as the number of rules increases (explained later in this chapter).
FPT	Fraud Prediction Threshold	The fraud prediction threshold is a set percentage. If a new claim comes in and its potential score of being fraudulent exceeds this threshold, the claim needs to be investigated.
SAW	Supervised Association Rule Weight	This percentage is the value that is added to the total score of whether a claim is fraudulent when a new claim matches a rule that confirms the fraudulent claim indicator. The SAW is also a benchmark percentage and decreases as the number of rules increases (explained later in this chapter).
ICFPS	Insurance Claim Fraud Prediction Score	This is the total suspicion factor that a claim is fraudulent, based on the supervised association rules that are triggered, unsupervised association rules that are broken and the probability determined by logistic regression.
FCI	Fraudulent Claim Indicator	The Boolean indicator on a claim that indicates whether a claim is fraudulent or not.

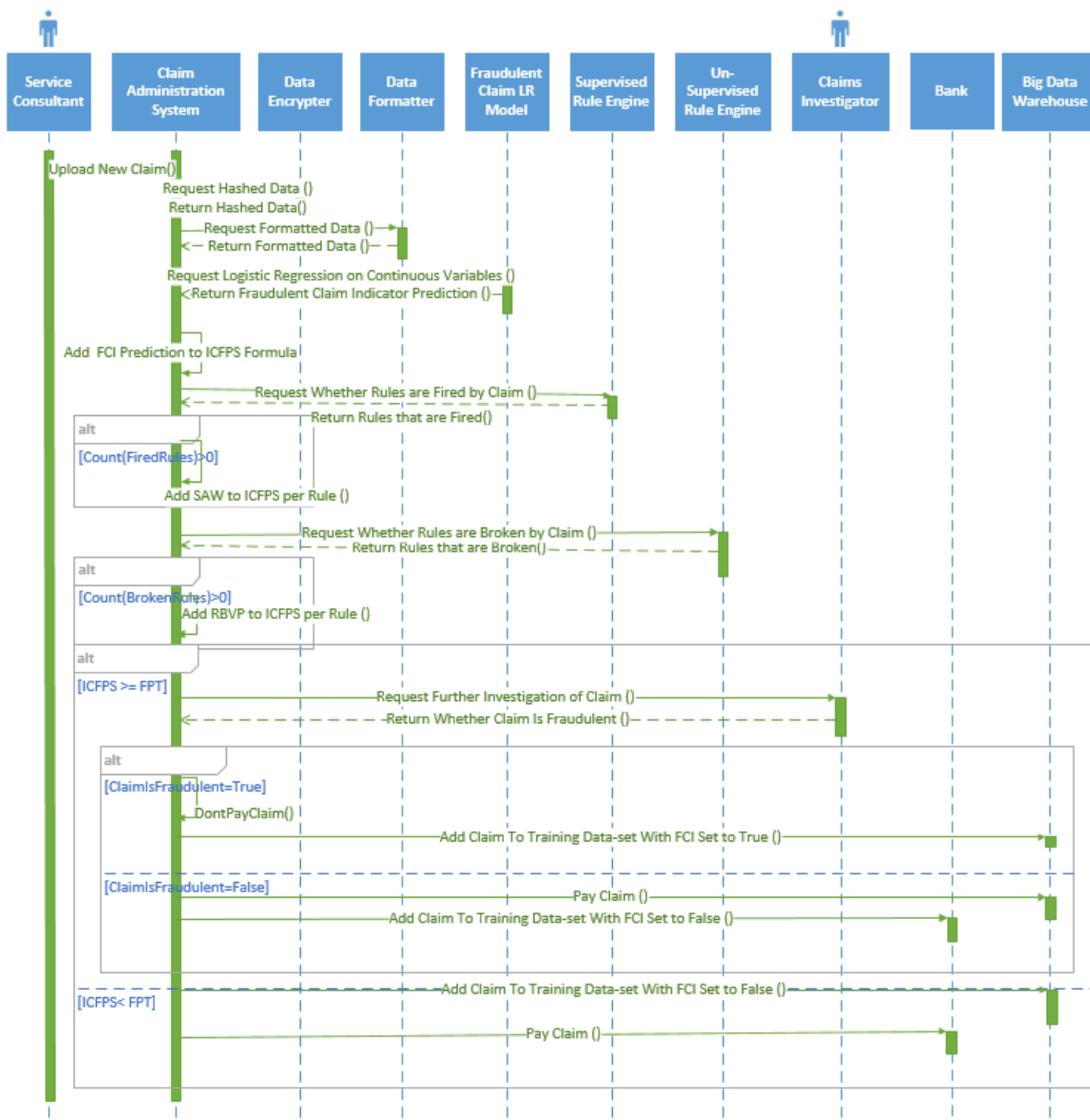


Figure 6.2: Sequence Diagram Showing the Submission of a Claim and Application of Knowledge to this Claim

Figure 6.2 shows a distinct application of these automatic recommendations. This process is described below:

1. A service consultant adds a claim to a claims management system. This claim is then encrypted so that anonymity of the claim can be safeguarded when adding the claim to the Big Data file system. The data needs to be consistent with the rules and as such, the data needs to be encrypted in the same way that training data has been encrypted.
2. Once this has been performed, the data is formatted correctly to work with the association rules and logistic regression model.
3. The claim is then passed through the logistic regression model where the fraudulent indicator can be predicted with a calculated confidence metric. The prob-

ability that the fraudulent claim indicator is true is then added to the Insurance Claim Fraud Prediction Score (ICFPS) formula (described further on in this chapter).

4. The claim is next passed through the supervised rule engine that indicates which rules contain the same characteristics as other claims that were fraudulent. For each of these rules that are triggered, the Supervised Association Rule Weight increases the ICFPS formula.
5. Subsequent to the supervised rules, the claim is further passed through the unsupervised rule engine. For every rule that is broken, the Rule Break Value Percentage increases the ICFPS formula.
6. If the ICFPS is greater than the Fraud Prediction Threshold, the claim needs to be sent to the claims investigation department where it can be investigated by a dedicated claims investigator. If the fraud investigator determines that the claim is indeed fraudulent, the claim will not be paid, and the claim can be added to the training data set with the fraudulent claim indicator set to true. If the fraud investigator determines that the claim is not fraudulent, the claim can be paid, and the claim is added to the training data set with the fraudulent claim indicator set to false.
7. If the ICFPS is less than the Fraud Prediction Threshold, the claim can be paid, and the claim can be added to the training data set with the fraudulent claim indicator set to false.

6.3.4 Derivation of Formula

From the above, it can be seen that a formula can be generated to depict the Insurance Claim Fraud Prediction Score (ICFPS). This can be derived from the process flow in the following way:

The Insurance Claim Fraud Prediction Score (ICFPS) needs to consider the number of supervised association rules that are triggered, the number of unsupervised rules that are broken, as well as the results of logistic regression.

The Insurance Claim Fraud Prediction Score is given by:

$$ICFPS = \frac{a}{x} + \frac{b}{y} + \frac{c}{z}$$

Where

a is the score that confirms a claim as fraudulent, using the supervised association

rules,

b is the score that confirms the claim as fraudulent, using the unsupervised association rules

c is the score that confirms a claim as fraudulent, using logistic regression.

Since it would be useful to have the ICFPS as a standard representation, it is stipulated that the ICFPS should be shown as a percentage between 0 and 100%.

Furthermore, x can be seen as the weighting of the supervised association rules score, y is the weighting of the unsupervised association rules score and z is the weighting of the logistic regression. The results of logistic regression are a probability that follows the basic statistical rule that the value must be between 0 and 1 (Levine and Stephan, 2009). Therefore, to standardise each score, it is stipulated that the score from the supervised and unsupervised association rules must be a value between 0 and 1 as well. It must therefore be further specified that $0 \leq a \leq 1$, $0 \leq b \leq 1$ and $0 \leq c \leq 1$.

The sum of the individual scores must be weighted such that $0 \leq ICFPS \leq 1$. From the maximum value of each score, we can then infer that $\frac{1}{x} + \frac{1}{y} + \frac{1}{z} = 1$ as a, b and c will have a maximum value of one each. If this is the case, then the total maximum score must be equal to one as well.

If we assume that all three machine learning techniques (supervised association rules, unsupervised association rules and logistic regression) are equally dependable in the prediction of fraud, then the formula can be further specified as:

$$ICFPS = \frac{a}{3} + \frac{b}{3} + \frac{c}{3}$$

In the event that the machine learning techniques mentioned are not equally reliable, the denominators can be weighted accordingly.

To generate percentage scores from the unsupervised and supervised rules, it is stipulated that each rule being triggered (SAW) is worth 15% and each rule being broken (RBVP) is worth 5%. This would be problematic, because if seven or more rules were broken, the percentage would exceed the 100% mark, which can be seen to be incorrect. It is suggested that instead of using 15% as a hard-coded percentage score for each rule that breaks, 15% should be used as the benchmark for one rule being broken, and as the number of rules increases, the percentage score increases and converges to 1 ($\lim_{x \rightarrow \infty} f(x) = 1$). An example of such a distribution can be seen through exponential decay towards a limiting value (negative exponential growth) (Marriott, 2013) that is based on the exponential function thought to be first explored by John Napier (1550–1617) (Engel and Nagel, 2000). This distribution fits the requirement,

as there comes a point where the number of rules that are broken no longer substantially increases the chance of a claim being fraudulent. This can be expressed with the formula $W = W_0 + A(1 - e^{-kt})$, which is shown in Figure 6.3 as derived from Marriott (2013). In this formula, W is the dependent variable, t is the independent variable, W_0 is the value where $t = 0$ and A is the value of the asymptote relative to W_0 (Marriott, 2013).

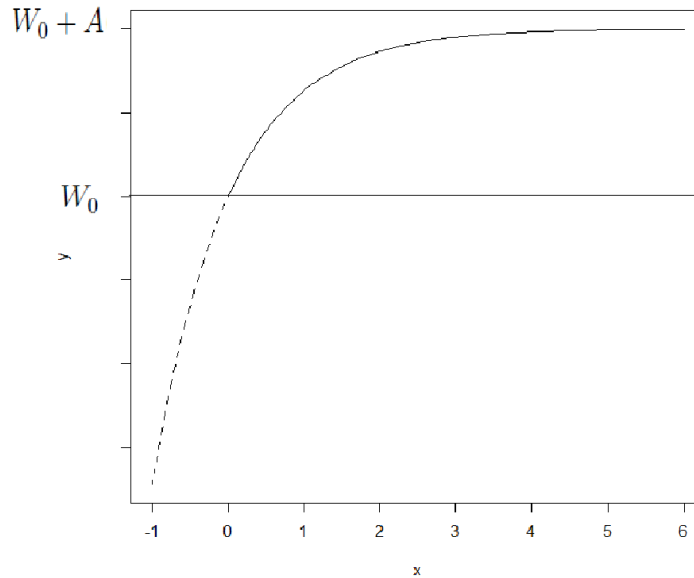


Figure 6.3: Chart Showing $W = W_0 + A(1 - e^{-kt})$ as derived from Marriott (2013)

If this formula is applied to the research in hand and it is stipulated that the formula of a is

$$a = W_0 + A(1 - e^{-kt})$$

then a is the score that a claim is fraudulent using the supervised association rules. If we specify that the percentage score starts at 0 and has a limit at 1, we can say that $W_0 = 0$ and $A = 1$, which will result in this formula being further specified as follows:

$$a = 1 - e^{-kt}$$

From the benchmark of one rule being triggered, k can be determined as follows:

$$0.15 = \frac{1 - e^{-k(1)}}{3}$$

$$0.450 = 1 - e^{-k(1)}$$

$$-0.550 = -e^{-k(1)}$$

$$0.550 = e^{-k(1)}$$

$$\text{Ln}(0.550) = \text{Ln}(e^{-k(1)})$$

$$\text{Ln}(0.550) = -k\text{Ln}(e)$$

$$-k = \text{Ln}(0.550)$$

$$k = 0.597837$$

Therefore, a can be specified as:

$$a = 1 - e^{-0.597837t}$$

This formula, which can also be more accurately written as $a = 1 - e^{\text{Ln}(0.550)t}$ so as to reduce a loss of precision, is illustrated in Figure 6.4.

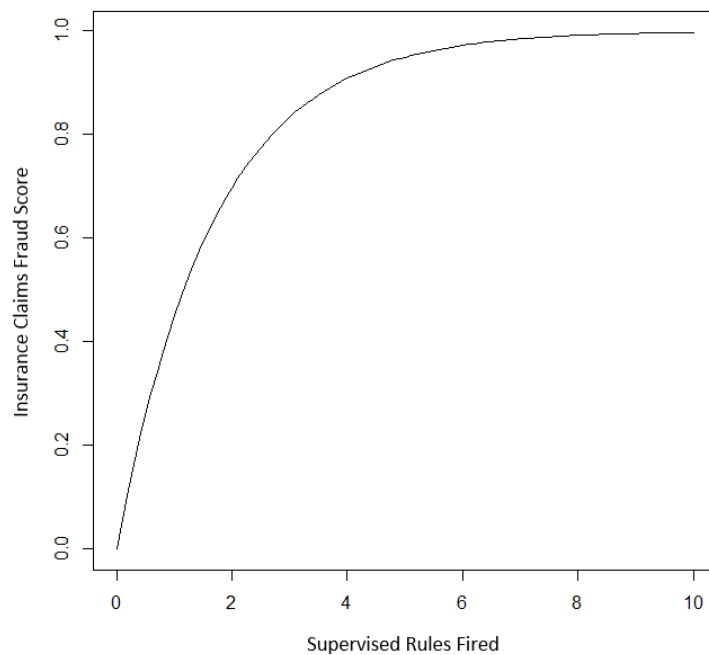


Figure 6.4: Chart showing $a = 1 - e^{-0.597837t}$ as derived from Marriott (2013)

The same formula can be used to determine b , if it is stipulated that the benchmark of an unsupervised rule being broken is 5%. Through the same process, the k value of the formula can be determined as follows:

$$b = W_0 + A(1 - e^{-kt})$$

$$b = 1 - e^{-kt}$$

$$0.05 = \frac{1 - e^{-k(1)}}{3}$$

$$0.150 = 1 - e^{-k(1)}$$

$$-0.850 = -e^{-k(1)}$$

$$0.850 = e^{-k(1)}$$

$$\text{Ln}(0.850) = \text{Ln}(e^{-k(1)})$$

$$\text{Ln}(0.850) = -k\text{Ln}(e)$$

$$-k = \text{Ln}(0.850)$$

$$k = 0.162519$$

Therefore, b can be specified as

$$b = 1 - e^{-0.162519t}$$

This formula can also be more accurately written as $b = 1 - e^{\text{Ln}(0.850)t}$, to reduce a loss of precision, as can be seen in Figure 6.5.

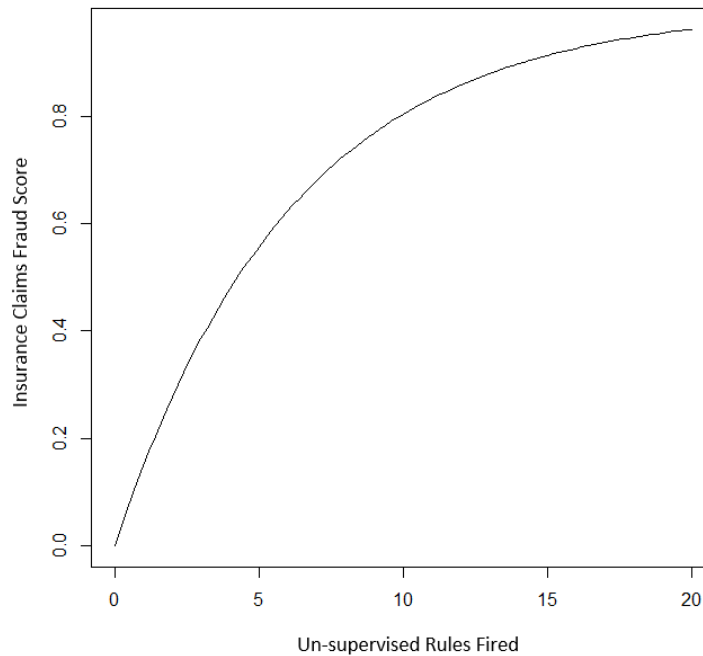


Figure 6.5: Chart showing $b = 1 - e^{-0.162519t}$ as derived from Marriott (2013)

For c , there is already a formula in play owing to the fact that logistic regression is being used (Kleinbaum and Klein, 2010). This formula is the logistic model formula

and not the logistic function formula – as the logistic model formula is seen as a number between 0 and 1. The logistic model is consequently described as follows:

$$c = \frac{1}{1 + e^{-(\alpha + \sum \beta_i X_i)}}$$

To understand this formula, it is important to note the following:

c is the dependent variable which estimates a likelihood that the claim imported is fraudulent,

X_i is an independent variable from the claims data set,

β_i and α are constants that are calculated each time the model is run. Hence, they are not specified in this derived formula (Peng et al., 2002).

The original formula of the score of insurance claims fraud can be further specified from a , b and c as follows:

$$ICFPS = \frac{a}{3} + \frac{b}{3} + \frac{c}{3}$$

$$ICFPS = \frac{1 - e^{Ln(0.550)s}}{3} + \frac{1 - e^{Ln(0.850)u}}{3} + \frac{1}{3(1 + e^{-(\alpha + \sum \beta_i X_i)})}$$

$$ICFPS = \frac{2 - e^{Ln(0.550)s} - e^{Ln(0.850)u}}{3} + \frac{1}{3(1 + e^{-(\alpha + \sum \beta_i X_i)})}$$

Here, s is the number of supervised machine learning rules that are triggered, and u is the number of unsupervised machine learning rules that are broken. ICFPS can therefore be seen to mark claims as needing investigation in Figure 6.2 when they are over the Fraud Prediction Threshold (FPT).

Based on the information and the formula described in this chapter, the ICFPM Model is described below.

6.3.5 Insurance Claims Fraud Prediction Model (ICFPM)

The ICFPM consists of three sub-models that are created to learn from the insurance claims data provided and that provide automatic recommendations from this data to evaluate whether new insurance claims are fraudulent or not. These three sub-models are the following:

1. Insurance Claims Fraud Apriori Association Rules Machine Learning Model (IFAMLM): The IFAMLM is a machine learning model that learns association rules from a claims data set. These association rules can be filtered into two sets of rules that include those with the consequent of confirming the fraudulent claim indicator as true and those having nothing to do with fraud.
2. Insurance Claims Fraud Logistic Regression Machine Learning Model (IFLMLM): The IFLMLM is a machine learning model that learns from an insurance claims data set using logistic regression. The IFLMLM outputs a probability of whether a claim is fraudulent or not.
3. Insurance Claims Fraud Prediction Score Model (ICFPSM): The ICFPSM takes the results of the IFAMLM and IFLMLM models and applies them to a new claim to give the claim a score of whether the claim should be investigated for fraud or not.

To train the IFAMLM and IFLMLM the following assumptions must be made:

Let $U = \{U_i; U \text{ is a Distributed Database of Insurance Claims}\}$

Let $D = \{d_1, d_2, \dots, d_n\}$

Where:

n is the number of claims used from the claims data set,

D is used to create the ICFPM model,

$$D = D^1 \cup D^2$$

Where:

D^1 = Randomly extracted set of claims used to train the model with 10% being fraudulent and 90% not fraudulent (Insurance Information Institute, 2017),

D^2 = Randomly extracted set of claims used to test the model with 10% being fraudulent and 90% not fraudulent (Insurance Information Institute, 2017).

Let $I = \{i_1, i_2, \dots, i_m\}$

Where:

I = a set of features of a claim called items,

m = the number of features of a claim,

$$i_i \in D \forall i_i = f(d_j, \dots, d_k)$$

Where:

$$j \geq 1,$$

$$k \leq n,$$

i_i can be a feature of a claim or the result of a function on the features of a claim.

6.3.5.1 The IFAMLM Sub-model

For the IFAMLM sub-model, the research provided by Agrawal et al. (1994) allows data scientists to generate a list of association rules. The model description that follows shows the association rules that the current research further separated into supervised and unsupervised rules. The supervised association rules would have “FraudulentClaimIndicator=1” as the consequent of the rule and the unsupervised rules would not include the “FraudulentClaimIndicator” in the consequent. The IFAMLM sub-model is described as follows:

Let $T = d_i$

Where:

$$T \subseteq I,$$

$$T \supseteq TID$$

Where:

$$TID = \text{Unique Identifier for } I,$$

$$X \subseteq T$$

Where:

$$X = \text{a set of some items in } I \text{ (Agrawal et al., 1994)}.$$

Let $ar = X \Rightarrow Y$

Where:

$ar =$ Association Rule,

$X =$ The antecedent of the rule,

$Y =$ The consequent of the rule,

$$X \subset I,$$

$$Y \subset I,$$

$$X \cap Y = \emptyset \text{ (Agrawal et al., 1994)}.$$

Let $AR = \{ar_1, ar_2, \dots, ar_q\}$

Where:

q = the number of association rules generated.

Let $AR^1 = \{x: x \in AR, \text{ the consequent of } x \text{ is that the claim is marked as fraudulent}\}$

Let $AR^2 = \{x: x \in AR, \text{ the consequent of } x \text{ is not fraud related}\}$

Where:

$$AR^1 \cap AR^2 = \emptyset.$$

6.3.5.2 The IFLMLM Sub-model

The IFLMLM sub-model is trained by taking the claims mentioned previously to weight a logistic regression model. The IFLMLM sub-model is described as follows:

Let $Z = \{z_1, z_2, \dots, z_k\}$

Where:

Z is a sub-set of I .

Let $f(r) = \frac{1}{1+e^{-r}}$ or $f(r) = \frac{1}{1+e^{-(\alpha + \sum \beta_i z_i)}}$ (Kleinbaum and Klein, 2010)

Where:

$f(r)$ = the probability that a claim is fraudulent,

$r = \alpha + \beta_1 z_1 + \beta_2 z_2 + \dots + \beta_k z_k$ (Kleinbaum and Klein, 2010)

Where:

α = to-be-determined intercept calculated from the training data set (Peng et al., 2002),

$\{\beta_1, \beta_2, \dots, \beta_k\}$ = logistic regression weighting coefficients from the training data set (Peng et al., 2002).

6.3.5.3 The ICFPSM sub-model

The ICFPSM sub-model uses the supervised association rules (AR^1) and unsupervised association rules (AR^2) from the IFMLM sub-model and the trained formula from the IFLMLM sub-model to create a score for a new claim. This score is used to determine whether the claim should be investigated. The ICFPSM sub-model is described as follows:

Let nc = New Claim

Let $a = |SR|$

Where:

$$SR = \{x: x \in AR^1, \text{ antecedent}(x) \subseteq nc\},$$

a = the number of supervised rules with the same features as the new claim.

Let $b = |USR|$

Where:

$$USR = \{x: x \in AR^2, \text{antecedent}(x) \subseteq nc, \text{consequent}(x) \notin nc\},$$

b = the number of unsupervised rules with the same antecedent features as the new claim, but with a different consequent feature.

Let ICFPS = Insurance Claim Fraud Prediction Score

Where:

$$ICFPS = \frac{2 - e^{Ln(0.550)a} - e^{Ln(0.850)b}}{3} + \frac{1}{3(1 + e^{-(\alpha + \sum \beta_i X_i)})} \text{ (as described previously)}$$

Let II = InvestigationIndicator

Where:

$$II = ICFPS \geq FPT$$

Where:

$$FPT = \text{Fraud Prediction Threshold}$$

Using these three sub-models hopefully provided insight into creating a model that can intelligently predict insurance claims fraud through Apriori association rules and logistic regression.

6.4 Model Maintenance

Although not many organisations maintain their predictive models long after they have been created (Eckerson, 2007), model maintenance is a key step towards ensuring that the predictive analytics model stays successful. It will result in better model reuse, increase performance and even minimise company overhead costs (Eckerson, 2007). Neither the rules generated nor the logistic regression model should remain static and both should adjust to new trends in claims data. The model maintenance process involves re-training the model on a regular basis so that the rules and regression models are updated when applied to new claims.

6.5 Summary

The four steps proposed in this chapter (see Figure 6.1) with their sub-steps can be better understood when illustrated by means of a diagram, and the generic diagram can be expanded to include the aforementioned steps and corresponding sub-steps – see Figure 6.6.

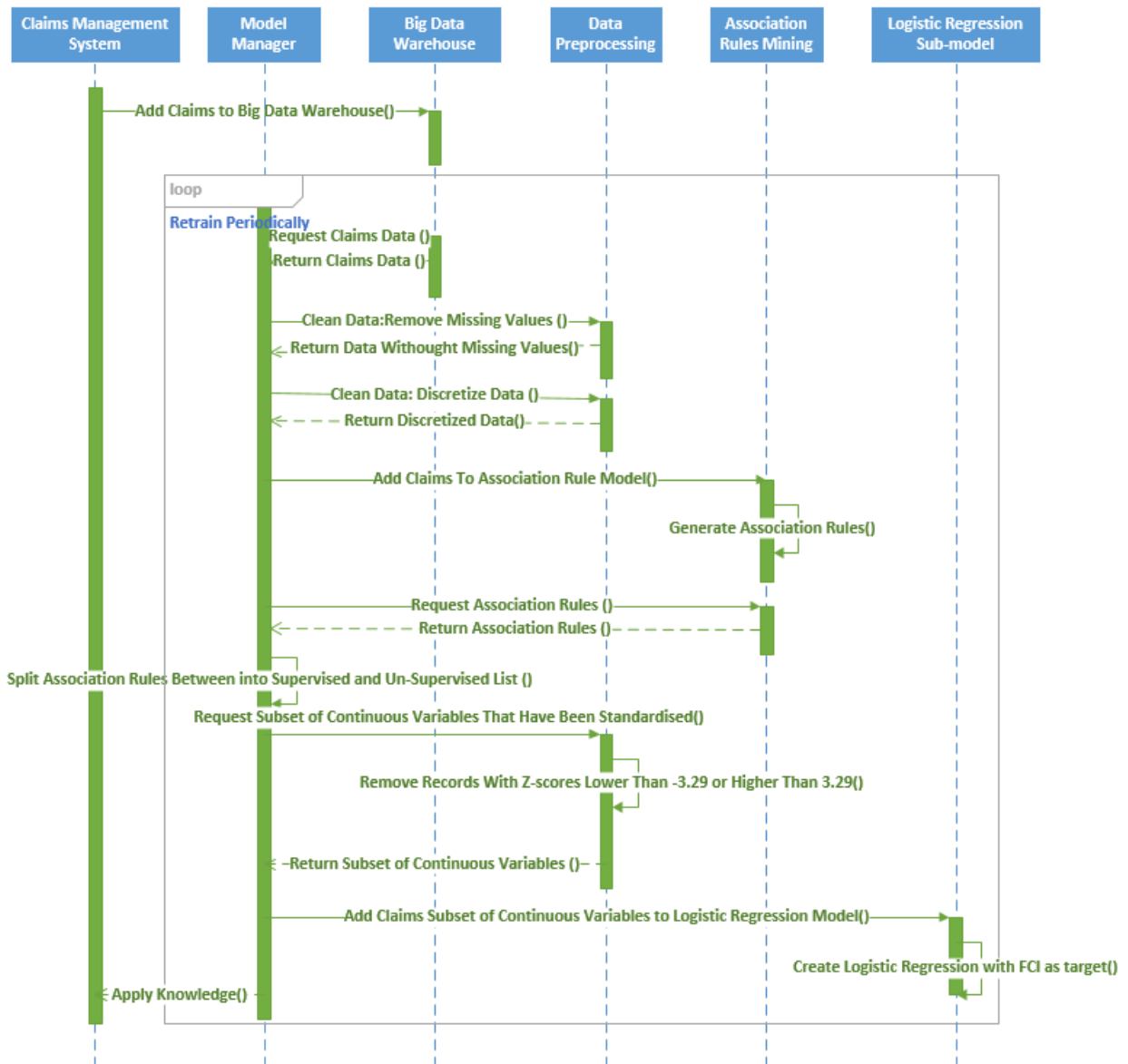


Figure 6.6: Sequence Diagram Showing Generation of a Solution that can Intelligently Predict Insurance Claims Fraud

6.6 Discussion

Chapter 6 described the steps required to create a system that can intelligently predict insurance claims fraud, namely data preparation, model creation, knowledge application and model maintenance. Each of these steps was described in more detail and design choices were motivated. Sequence diagrams were shown to gain an understanding of how all these steps fit together. A model definition was also put forward to describe how to intelligently predict fraud with logistic regression and association rules.

To facilitate this process, an understanding of required technologies needs to be fostered. This will be achieved in the following chapter with a description of a high-level architecture.

7 Architecture

To facilitate the process mentioned in Chapter 6, the architecture of the system needs to be fit for its purpose and meet the needs of a solution that can intelligently predict insurance claims fraud. Chapter 7 describes the possible design choices when creating such a solution that uses Big Data frameworks and data science platforms. The research in hand also puts forth a suggested design that has been tested and is known to work. The chapter uses previously determined software components such as the choice of data science platform and Big Data framework.

An architecture is described as the “art or science of building or constructing edifices of any kind for human use” (Oxford English Dictionary Online, 2017b). If architecture is refined to systems architecture, it is described as the “conceptual structure and overall logical organisation of a computer or computer-based system from the point of view of its use or design; a particular realisation of this” (Oxford English Dictionary Online, 2017a). A software architecture has also been described as “the hierarchical composition of computational and structural patterns, which we refine using lower-level design patterns” (Catanzaro et al., 2010). Hence, the architecture shown in this chapter will contain the conceptual structure of the system that can intelligently predict insurance claims fraud. It will however not be limited to software only.

To structure this chapter, existing literature contributions on what architectures must contain were used. Len et al. (2003) maintain that software architectures must include the structure of a system with its software components and the relationships between those components. Meier et al. (2009) in turn argue that an architecture must satisfy the technical, operational and quality requirements of a solution, while Bachmann et al. (2000) explore the notion that architectures should have layers with elements or components, relationships and properties. Based on these three literature contributions, the researcher derived the structure of this chapter and consolidated the approaches adopted in them.

Chapter 7 starts off by describing the operational, technical and quality requirements of the system. It must be noted, however, that the requirements are not the same as the requirements in the framework chapter; instead, here they are requirements of what is necessary to run the software itself and do not concern the functionality of the system. The chapter proceeds to describe the layers of the architecture and finally uses the requirements and layers to show the components and relationships of

the architecture in a structure. For this research, the focus is more on the necessary technical requirements to get the system running, and not the most adequate usability.

7.1 Operational Requirements

The operational requirements of the system have been derived from the software goals of the system (Van Lamsweerde, 2003). Hence, this research describes the necessary software requirements with relation to the goals of each component of the system below.

- **REQ 1:** The system must use a distributed file system, namely Hadoop (as determined in previous chapters) to accommodate Big Data. This will be done by using the MapR distribution of Hadoop.
- **REQ 2:** As it has previously been determined that Hive is a good way to store and process data in Hadoop, connecting to Hive and running HQL queries through R is necessary. A solution to this problem is RHive (Yang, 2013). RHive allows data scientists to run distributed computing algorithms from R through HQL (Lu and Zheng, 2013).
- **REQ 3:** The system must use R (as determined in previous chapters) to comply with the data science requirements. Alternative distributions of R such as Microsoft R Server (Revolution R) and Microsoft R Open (Salloum et al., 2016) have been created to result in the analytics performed by R, work with bigger data sets and are more scalable. Hence, the system will use Microsoft R Server.
- **REQ 4:** Since R is a scripting language and not immediately usable by other systems, it needs to be extended to have such functionality. One such extension is DeployR (Maartens, 2017a). DeployR allows the functionality of R to be used by other systems – be they mobile, web or desktop – by creating a web service that can be called by other systems to run R scripts and get results (Maartens, 2017a). This would result in claims management systems being able to access R and acquire recommendations and rules from R.
- **REQ 5:** The system must be incorporated into existing insurance claims management systems. The type of insurance claims management system does not have a direct impact on this research, and hence the researcher chose a Windows Model View Controller (MVC) architecture to test this research. The Windows MVC architecture would be replaced with whichever technology the existing insurance claims management system uses.
- **REQ 6:** It was determined in previous chapters that OpenRefine can be used for data cleaning. It is for this reason that OpenRefine must sit at the nexus between the insurance claims management system and Big Data file system.

7.2 Technical Requirements

A technical requirement is a non-functional or environmental requirement that involves the necessary features of the system to perform adequately (Tappenden et al., 2009). The current research therefore describes the technical requirements that must be met for the system to perform adequately.

Because it has been determined that the distributed file system will run on Hadoop, some of the minimum technical requirements of the version that was used for this research will be described below. Although all the minimum technical requirements for the version that was used can be viewed on <http://doc.mapr.com/display/MapR/Preparing+Each+Node> (MapR, 2017), the researcher aims to show the most important requirements relating to this research. Hence, minimum requirements like the operating system have been excluded and should be upscaled with an increase in data and the need for higher computational power.

The technical requirements are requirements that were met at the time of this research and they are non-exhaustive.

- **REQ 7:** Each node in the cluster must have a 64-bit processor. For this research, it can be seen to be important as it focuses on large data sets and complex calculations; hence the increased computational capacity that one gains from a 64-bit processor is necessary. Although one might not need a 64-bit processor when a claims data set is small, it will become increasingly necessary to have a 64-bit processor when the claims data set increases to a less manageable size.
- **REQ 8:** Each node in the cluster must have at least 8 GB of RAM but even more in a production environment. This requirement is once again only necessary as a claims data set increases in size and the complexity of calculations increases. Although one would have many nodes in a cluster, having failing nodes due to low memory would be an unnecessary fall-back.
- **REQ 9:** Each node in the cluster should have at least three physical drives with a minimum of 10 GB of space on the partition containing the operating system; at least 8 GB for the MapR file system; at least 10 GB in the “/tmp” directory; 128 GB of free space in the “/opt” directory, and at least 24 GB of swap space. This storage requirement is important when the claims data set grows rapidly, as one would not want to have to increase the number of nodes regularly, owing to the fact that a single node does not have much storage.

Due to the fact that DeployR will be used, the technical requirements of DeployR need to be mentioned. The full list of requirements for the version that was used were described by Maartens (2017c) and can be accessed at <https://docs.microsoft.com/enus/machine-learning-server/deployr/deployr-installing-configuring>. From the requirements mentioned in that article, the requirements most relating to this research are described

below:

- **REQ 10:** For the server where DeployR is loaded, a minimum processor speed of 3.0 GHz is recommended. Although most of the computation will be done by the data science tool (R) and Hadoop, DeployR cannot become a bottleneck. Owing to the fact that a claims management system would access DeployR to get real-time recommendations or to access rules about claims, it would be highly problematic for a service consultant to be slowed down by a web service not performing computations and only providing recommendations.
- **REQ 11:** The server requires at least 4 GB of RAM. This may seem redundant, but owing to the fact that DeployR requires other software such as Apache Tomcat, MongoDB and Java, it is important to ensure that the server has enough memory not to slow down automatic recommendations of whether claims are fraudulent or not. It is important to mention that Apache Tomcat, Java and MongoDB are software requirements to facilitate the usage of DeployR. MongoDB would not be storing any of the insurance claims data set.
- **REQ 12:** It has been mentioned that internet access is a requirement for DeployR. Although most of the components in this solution would require internet access, determining which components should have internet access to outside the organisation would require critical analysis and would depend on whether each requirement would be accessed by external parties. For example, it would be better to only allow the server running the claims management system to access DeployR, as it could be considered more secure to not expose the service to everyone.

As the system will use Microsoft R Server, the following requirements have been derived from Steen (2016). The full list of requirements can be found at <https://docs.microsoft.com/en-us/machine-learning-server/install/r-server-installlinux-server-805> but once again, only the requirements that add value to this research are mentioned below. For the servers running Microsoft R Server software, the following requirements apply:

- **REQ 13:** The server requires a minimum of 2 GB of RAM but 8GB or more is recommended. Although this stipulates a minimum of 2GB of RAM, the researcher suggests that having as much RAM as possible, without incurring unnecessary expense, is required in the data science platform so as to not result in slow responses when trying to make recommendations about fraudulent claims.
- **REQ 14:** The server requires at least 500 GB hard drive space. Although this requirement is perfectly valid, the claims data set would be kept in the distributed file system. As such, this requirement would relate more to the R software itself and the data currently being worked on.
- **REQ 15:** The server requires a 64-bit processor. As with the machine learning algorithms, this is once again an important requirement to mention. If there is an

increase in complexity, one would not want to be restricted by a 32-bit processor. Insurance claims are primarily financial and hence the complexity of numbers could increase.

Because OpenRefine will be used for data cleaning, it is also necessary to mention the technical requirements of this data cleaning tool. The technical requirements for OpenRefine are vague (Stephens, 2017), probably due to its lightweight nature, and as such the technical requirements of this data cleaning tool will not be mentioned.

7.3 Quality Requirements

According to Meier et al. (2009), quality requirements include manageability of a system, security of a system and performance of a system.

7.3.1 Manageability

The manageability of a system refers to how easy is it to maintain and monitor the system (Young, 2007). The two requirements explored below relate to monitoring and maintaining the aforementioned software:

- **REQ 16:** It is necessary to have an easier way of maintaining the Hadoop nodes in the cluster. Because it has already been determined that MapR is the Hadoop distribution of choice, the MapR Control System can be used to manage the cluster (MapR, 2017). This should ensure that the nodes in the cluster are functioning correctly and if problems arise (such as the claims data set becoming too large), this can be noticed on the MapR Control System.
- **REQ 17:** Because R scripts are not directly usable by systems, and because DeployR will be used to turn these scripts and their results into a web service, there needs to be a way to manage this process. DeployR offers a Repository Manager that allows data scientists to manage the scripts that are to be used by other systems, and in this case, by the insurance claims management system (Maartens, 2017b). Thus, new versions of the Apriori association rules and logistic regression scripts that are applied to the claims data set can be updated and modified through a user interface.

7.3.2 Security

The security of a system refers to the fact that software must continue to perform correctly when it is maliciously attacked (McGraw, 2004). Since a critical analysis of the security of such a system is beyond the scope of this research, only basic security requirements are shown below.

- **REQ 18:** It was previously mentioned that policyholder information will be hashed for privacy reasons. It is also necessary to protect the data as a whole as it is insurers' and brokers' data. Because the data is kept in HDFS in the MapR version of Hadoop, one can use the security provided for by MapR. MapR has encryption, authorisation and authentication security features to ensure that data is secure (MapR, 2017).
- **REQ 19:** Ensuring that DeployR does not allow a user direct access to the data is also important. Hence, the security features of DeployR should be used. DeployR has basic authentication as a feature, as well as HTTPS and SSL support (Maartens, 2017d).

7.3.3 Performance

The performance of the system was addressed by the use of a cluster computing framework and the introduction of minimum hardware requirements.

7.4 Layers

According to Bachmann et al. (2000), layers contain certain sets of functionality that can be grouped together. These layers partition the software and can be added and removed where necessary. The layers of the proposed solution that can intelligently predict insurance claims fraud are described below.

- **LAYER 1: Presentation Layer** - the layer of the architecture that a service consultant sees as processing claims through an insurance claims management system.
- **LAYER 2: Application Layer** - the layer of the architecture that uses the information gained from results delivered by data science tools to predict whether a new claim processed by a service consultant needs to be flagged as possibly fraudulent or not.
- **LAYER 3: Data Layer** - the layer of the architecture in the insurance claims management system that houses the operational database.
- **LAYER 4: Data Science Layer** - the layer of the architecture that contains the machine learning models (ICFPM, IFAMLM, IFLMLM and ICFPSM) that are used to predict fraud.
- **LAYER 5: Big Data Layer** - the layer of the architecture that contains all the historical claims data sets.

7.5 Structure of the System

To explain the structure of the architecture, the software and hardware components of the system need to be shown. To model the software and the hardware of the system using Unified Modeling Language (UML), a component diagram and a deployment diagram were used respectively (Kobryn, 2000). They are shown in Figure 7.2 and Figure 7.3 with a software view of the architecture and a hardware view of the architecture. Figure 7.2 and Figure 7.3 were based on the requirements described in this chapter. The diagram shown in Figure 7.1 depicts the structure of the system that will be populated with the necessary software and hardware for the software (see Figure 7.2) and hardware (see Figure 7.3) view.

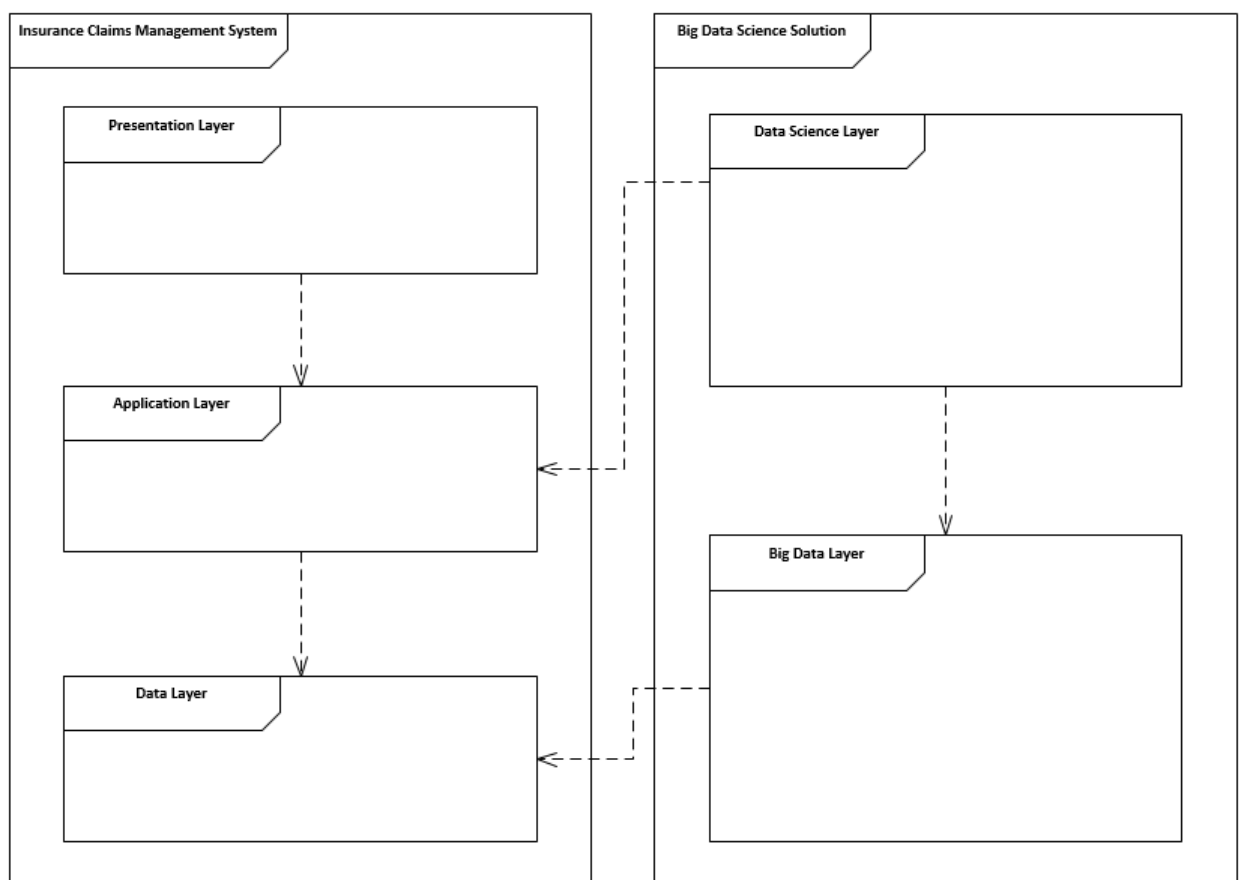


Figure 7.1: The Structure of the System to be Populated with Software and Hardware

7.5.1 Software View

The best approach towards the intelligent prediction of insurance claims fraud as taken in this research is depicted in the architectural diagram shown in Figure 7.2. The diagram shows the Insurance Claims Management System on the left, with the Big Data and data science components on the right. Parts of the insurance claims management system such as the ASP.Net view and C# controller are only shown because they were

necessary to test the solution and not because they were the best choice for a system that can intelligently predict insurance claims fraud. The left side of the diagram shows a generic insurance claims management system that would use a commonly described 3-tier architecture, namely MVC (Model-View-Controller) (Wang et al., 2009).

The ASP.Net View and C# controller would be replaced with whichever technology the existing insurance claims management system uses. For the purposes of this research, it was assumed that the ASP.Net View and C# controller would be used. The architecture, using a Big Data framework and data science platform, would include R as the data science platform of choice, and R would use RHive to run HQL queries. These HQL queries are relayed to MapReduce jobs in Hadoop, based on the data in HDFS. The data in HDFS is imported using Sqoop and HDFS commands. This import is from the Insurance Claims Management System SQL Database and External CSV Data.

The C# controller would be adding new policyholder and claims data to the SQL Database. R scripts would be run from web service calls through the DeployR API. This would result in the Insurance Claims Fraud Logistic Regression Machine Learning Model (IFLMLM) and rules generated from the Insurance Claims Fraud Apriori Association Rules Machine Learning Model (IFAMLM) as an output. The rules can be converted to XML and be added to an XML Rules Engine that can be understood by the C# controller. These rules could then be interpreted by the C# Controller and the logistic regression model can be applied as a Logistic Regression Predictor used to predict fraudulent claims. This way, when new claims are added through the ASP.Net view, they can be predicted as fraudulent or not fraudulent, based on the XML Rules Engine and Logistic Regression Predictor.

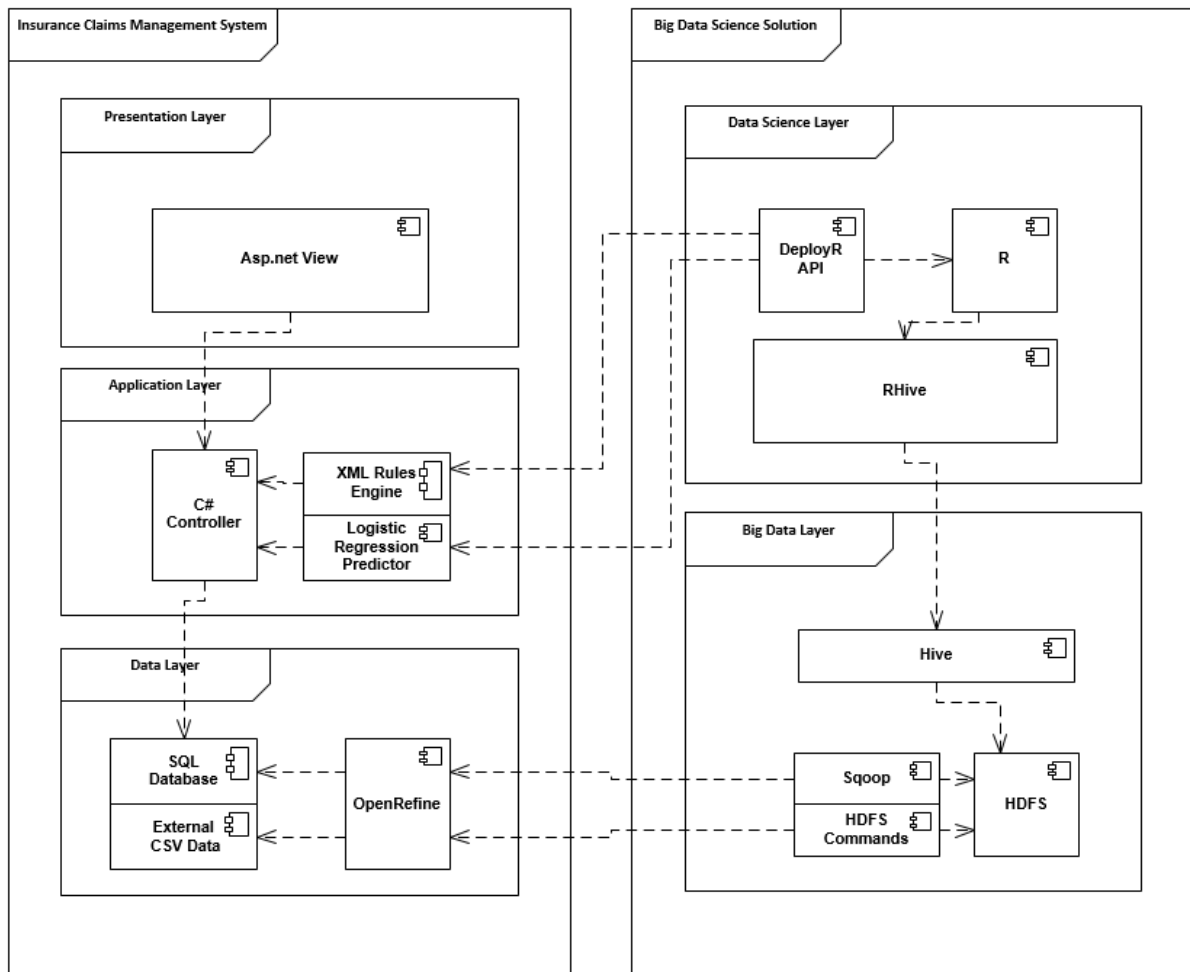


Figure 7.2: Software Architecture of the Solution that Intelligently Predicts Insurance Claims Fraud

7.5.2 Hardware View

A view of what the hardware would look like have been derived from the technical requirements mentioned in this chapter. This view is shown in Figure 7.3. The diagram shows existing infrastructure on the left that forms part of service consultants processing claims through an insurance claims management system. The diagram shows that a computer is used to access a claims management system (cms) website that is hosted on a Web Server that accesses an operational database (claimsManagementSystemDB) on a Windows Server. The right-hand side shows the hardware that would be necessary for the data science software and Big Data file system. This would include many servers that run Centos with a version of MapR Hadoop on them. The RServer Red Hat Server would contain DeployR and Microsoft R Server, which could access the managed cluster. The technical specifications of the servers shown in this diagram have already been described in this chapter and hence will not be repeated. The operating systems used were derived from the requirements of the software but can be replaced by any supported operating system.

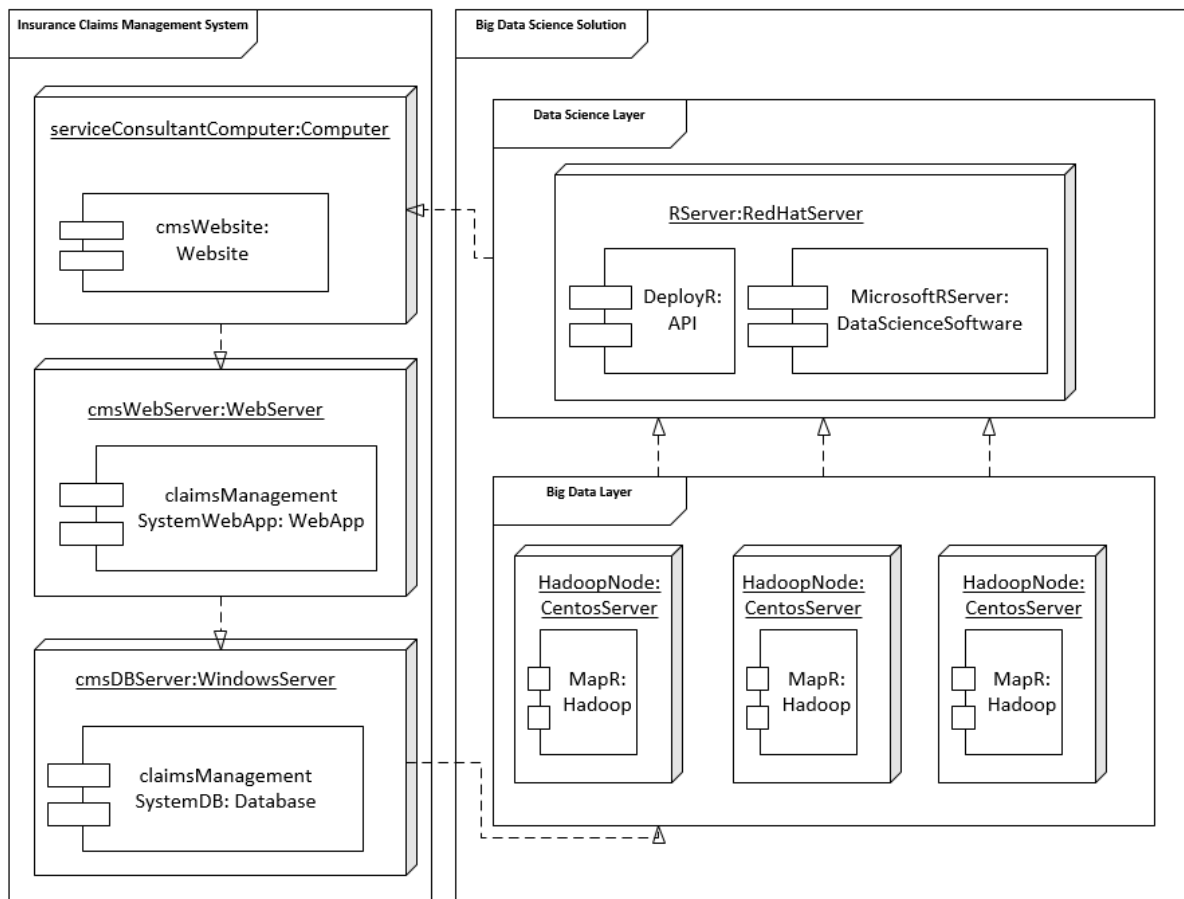


Figure 7.3: Hardware Architecture of the Solution that Intelligently Predicts Insurance Claims Fraud

7.6 Discussion

This chapter described the possible design choices when creating a solution that is intended to intelligently predict insurance claims fraud. This was achieved by describing the operational, technical and quality requirements, as well as the layers that were used to derive a software and hardware architecture.

From this chapter and the previous chapters, an understanding should be fostered of how to create a solution that can intelligently predict insurance claims fraud. The design, processes and architecture that have been implemented need to be tested, and therefore the following chapter validates the prototype developed throughout the preceding chapters.

8 Constructing and Validating The Prototype

The previous three chapters detailed a proposed framework, model and architecture for intelligent insurance claims fraud prediction. The proposed framework, model and architecture were used to construct a prototype, and Chapter 8 now focuses on constructing this prototype, as well as on the tests that were performed to validate it.

This chapter is split into two main sections that deal with the construction of the prototype and its validation. The two sections are introduced below.

8.0.1 Construction of the Prototype

To structure the first section, a reminder of the software components of the prototype and the model definition can be beneficial to show what was built for the prototype to be tested. The diagram shown in Figure 8.1 shows that to construct the prototype, an insurance claims management system was connected to the trained Insurance Claims Fraud Prediction Model (ICFPM).

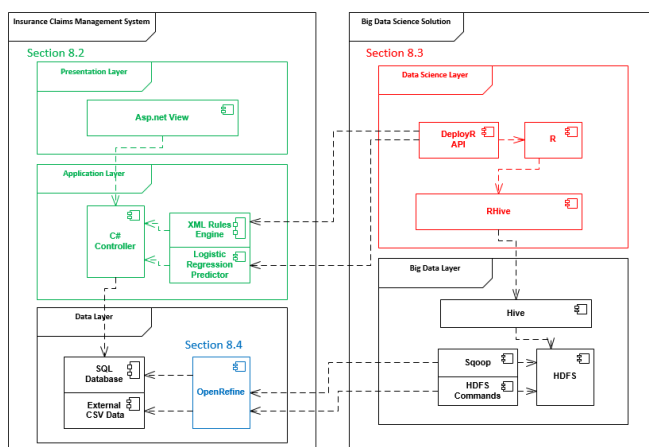


Figure 8.1: Software Architecture of the Solution that Intelligently Predicts Insurance Claims Fraud

The full model definition was earlier defined in Section 6.3.5, but a summary of the model can be beneficial to understand Chapter 8. The Insurance Claims Fraud Prediction Model (ICFPM) includes the Insurance Claims Fraud Apriori Association Rules

Machine Learning Model (IFAMLM), the Insurance Claims Fraud Logistic Regression Machine Learning Model (IFLMLM) and the Insurance Claims Fraud Prediction Score Model (ICFPSM). These three models are used to calculate an Insurance Claim Fraud Prediction Score (ICFPS) for each new claim inputted into the system. The formula for ICFPS is:

$$\text{ICFPS} = \frac{2 - e^{Ln(0.550)a} - e^{Ln(0.850)b}}{3} + \frac{1}{3(1 + e^{-(\alpha + \sum \beta_i X_i)})}$$

The first fraction considers the Apriori association rules where a is the number of supervised rules with the same features as the new claim, and b is the number of unsupervised rules with the same antecedent features as the new claim, but with a different consequent feature. The second fraction considers the result of the logistic regression formula.

8.0.2 Validation of the Prototype

The discussion that follows describes the testing of this prototype. Although statistical accuracy testing should be sufficient, scenario-based testing was used as well to provide the reader a clearer understanding as to how the ICFPM model would work. Chapter 8 describes three scenarios that would flag a claims transaction as fraudulent (see Section 8.2.2). Scenario-based testing was also put forward to determine whether PPDM worked with this prototype. Lastly, the architecture is tested to determine whether it can cater for large numbers of insurance claims.

The two sections below reference sub-folders where components of the prototype are kept. To access these components, navigate to <https://tinyurl.com/UPInsurClaimsFraud>. You are regularly referred to sub-folders in Chapter 8 where components of the prototype can be found. These sub-folders are referred to in italics and within parenthesis.

8.1 Constructing the Prototype

To construct the prototype, a system was created to act as an insurance claims management system. This was then connected to the ICFPM model which would predict fraud. The sub-sections that follow describe the generation of the insurance claims management system prototype, the generation of test data, the cleaning of test data and how to train the ICFPM model with this data that has been filtered and transformed to work with the model.

8.1.1 Generation of the Prototype: Insurance Claims Management System

To do the scenario testing, a Model View Controller (MVC) application was created. This application allowed a user to capture a new claim or add a list of claims. A large amount of information was required during claim submission, and the web page view showed four different tabs of data to be completed before a claim could be submitted: policyholder information (Figure 8.2); policy information (Figure 8.3); claims information (Figure 8.4); third-party information (Figure 8.5). The website also included a screen where one could set the maintenance variables of the system (Figure 8.6).

POLICYHOLDER INFORMATION	POLICY INFORMATION	CLAIMS INFORMATION	THIRD-PARTY INFORMATION
Age:	30.00		
Date of birth:	01/11/1990		
Gender:	Male		
Insured name:	test		
Insured surname:	Kenyon		
Marital status:	Married		
Policyholder's area:	Bryanston		
Policyholder's city:	Johannesburg		
Policyholder's postal code:	2158		
Policyholder's province:	Gauteng		
Policyholder's telephone number:	011 465 8958		
Policyholder's street:	Ballyclare Drive		

← POLICY INFORMATION

Figure 8.2: Adding Policyholder Information to the Prototype Insurance Claims Management System

POLICYHOLDER INFORMATION	POLICY INFORMATION	CLAIMS INFORMATION	THIRD-PARTY INFORMATION
Total policies revenue:	1,600.00		
Policy start date:	01/11/2011		
Policy end date:	01/11/2012		
Agent:	M Naidoo		
Broker:	AB Brokers		
Insurer:	Insurer A		
Sum insured:	100,000.00		
Insurance coverage type:	Fire and theft		
← CLAIMS INFORMATION			


Figure 8.3: Adding Policy Information to the Prototype Insurance Claims Management System

POLICYHOLDER INFORMATION	POLICY INFORMATION	CLAIMS INFORMATION	THIRD-PARTY INFORMATION
Claim payment amount:	15,000.00		
Postal code:	2158		
Province:	Gauteng		
Area:	Bryanston		
City:	Johannesburg		
Date of claim:	01/11/2011		
Date of loss:	01/11/2011		
Kind of loss:	Theft		
Probable amount of entire loss:	30,000.00		
Amount claimed:	30,000.00		
Excess paid:	1,000.00		
Payment account number:	1542512556		
Branch code:	45126		
Assessor's unique identifier:	NA		
Title holder's name:	David		
Title holder's surname:	Kenyon		
Short description:	test		
Long description:	test		
Claim service provider:	AZR Motors		
← THIRD-PARTY INFORMATION			

Figure 8.4: Adding Claims Information to the Prototype Insurance Claims Management System

POLICYHOLDER INFORMATION	POLICY INFORMATION	CLAIMS INFORMATION	THIRD-PARTY INFORMATION
Other party's name:		<input type="text" value="NA"/>	
Other party's surname:		<input type="text" value="NA"/>	
Other party's damage:		<input type="text" value="0.00"/>	
Other party's insurer:		<input type="text" value="NA"/>	
Other party's street:		<input type="text" value="NA"/>	
Other party's province:		<input type="text" value="NA"/>	
Other party's city:		<input type="text" value="NA"/>	
Other party's area:		<input type="text" value="NA"/>	
Other party's postal code:		<input type="text" value="test"/>	
<input type="button" value="SUBMIT CLAIM"/>			

Figure 8.5: Adding Third-Party Information to the Prototype Insurance Claims Management System



**Insurance Claims
Fraud Finder**

[Home](#) [Add Claims](#) [View Claims](#)

Maintenance

Rule Break Value Percentage:	<input type="text" value="0 %"/>
Fraud Percentage Threshold:	<input type="text" value="0 %"/>
Supervised Association Rule Weight:	<input type="text" value="0 %"/>

Figure 8.6: Web Page to set the Maintenance Variables

Whenever a claim was submitted, it was evaluated as being fraudulent or not. Whether or not it exceeded the fraud percentage threshold would determine whether the dialogue shown in Figure 8.7 would appear (*sub-folder:4. MVC Prototype Code | Views*).

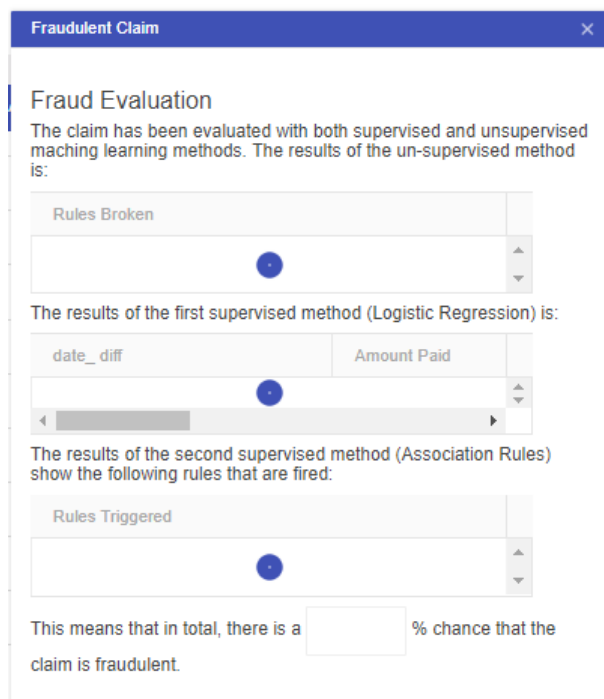


Figure 8.7: Dialogue Shown When Claim is Flagged as Potentially Fraudulent

8.1.2 Generation of the Test Data

At the start of this research, the researcher intended to gain access to real-world insurance data to develop the model and test it. It was agreed with an A-rated insurer in South Africa that their data would be used to create and test the model ("General Manager: Risk and Technology", 2016). This plan had to be abandoned owing to ethical and legislative considerations, as it is unethical to mine policyholders' data without their consent. For this reason, a combination of both real-world insurance data and fabricated policyholder data was used, and the data set that was generated consisted of 100 000 claims. An insurance company in South Africa on average receives approximately 130 000 insurance claims in a year (BusinessTech, 2015) and to ensure the accuracy of a predictive analytics model, between 2 000 and 60 000 claims would be needed (Liao and Zhu, 2014) (*sub-folder:1. Data-Sets\5. Test Data*).

8.1.2.1 Real-world Data

Transactional data – i.e. data involving the debit order collection of insurance premiums (premium collection) and claims payment data – was taken from a real-world system so that the distribution of the data would be correct. The premium collection data included a unique policyholder identifier that was linked to the claims payment unique policyholder identifier. This unique policyholder identifier was replaced with a fabricated, random numeric identifier, and this consequently resulted in a data set with a fabricated policyholder identifier, a total policy revenue, claims payment amount,

policy start date and claim payment date. As this did not include any personal information, it could not be seen as a breach of the PoPI Act (Luck, 2014). A policyholder could not be identified from this data, unless the person with the data had access to the original database and could determine the period(s) for which the extract had been run. The total policy revenue and claims payment amount were used to estimate other figures such as the sum insured of the policy (*sub-folder:1. Data-Sets\1. Real-world Data*).

8.1.2.2 Fabricated Data

After this, the fabricated unique policyholder identifier was randomly allocated to fabricated claims. These fabricated claims included fabricated policyholder information, claims information, policy information and third-party information. The fabricated policyholder information included information such as the name, surname and addresses of the policyholder. The policy information included information such as who the broker was, who the insurer was, and the policy end date. The claims information included data such as the date of the claim, and the loss date. Third-party information included data such as the name and surname of the other party. It could clearly be noticed from some of the fabricated data mentioned that this information was personal, and using it without the policyholders' consent would be a breach of PoPIA (*sub-folder:1. Data-Sets\2. Fabricated Data*).

8.1.2.3 Calculated Fields

Two calculated fields were appended to the training data set. The difference between the date of a claim and the start date of the policy was added as `DateDifference`. This was seen to be important, as it resulted in a numeric representation of how long it took a policyholder to claim, instead of just dates. Another field, that was added – `ClaimedOverPaid` – was a ratio of how much was claimed over how much premium was paid by the policyholder. This was meant to give a more standard representation of a claim amount, as policyholders have widely varying claim amounts. This representation of the ratio of total losses (claims) vs the total premium received is regularly used by insurance companies and is known as the loss ratio (Simon-Tuval et al., 2015) (*sub-folder:1. Data-Sets\3. Calculated Fields*).

For a description of the full data set, excluding the calculated fields, one can refer back to Chapter 2. The fields described were used for the original training data set. The fields were filtered to train the Insurance Claims Fraud Apriori Association Rules Machine Learning Model (IFAMLM) as well as the Insurance Claims Fraud Logistic Regression Machine Learning Model (IFLMLM). Since these are two different machine learning algorithms with different features, the fields used are not the same. Therefore the filtering of the data before training the ICFPM is described later on in the chapter.

8.1.3 Cleaning the Test Data

8.1.3.1 Initial Claims Management System Cleaning

To clean the test data in a live environment, the data first had to pass through OpenRefine to be standardised. This was done with the source data before it was hashed for PPDM. The data was subsequently hashed and exported from the source system. In the case of the test that was performed, the data was standardised using OpenRefine and subsequently all personal information and personal identifiers were hashed (*sub-folder:1. Data-Sets\4. Hashed Data*).

8.1.3.2 Data Quality Cleaning

Once the hashing was complete, the data was checked for data quality and records that had poor data quality were removed from the training data set. Steps to clean the data are described below. Because OpenRefine API calls can be lengthy, only one example is discussed in this chapter and more examples are presented in the appendix (*sub-folder:3. Data Cleaning*).

- Empty lines were removed from the data. This was done using the OpenRefine API call shown in Appendix E.

In addition to removing empty lines, further data quality checks were performed. A check was performed to ensure that fields were of the correct data type:

- DateOfClaim, PolicyStartDate and PolicyEndDate were checked to ensure that they were of the date type. An example of the OpenRefine API call for DateOfClaim is shown in Appendix F. This was repeated for PolicyStartDate and PolicyEndDate.
- SumInsured, Age, ClaimAmount, AmountPaid were all checked to see if they were numeric. An example of the OpenRefine API call for AmountPaid is shown in Appendix G. This was repeated for SumInsured, Age and ClaimAmount.

Further data quality checks were performed to ensure that the data followed basic business rules. These included the following:

- A check to see that the PolicyStartDate was before the PolicyEndDate. The OpenRefine API call for this is shown below.

```
1 [
2   {
3     "op": "core/row-removal",
4     "description": "Remove rows",
5     "engineConfig": {
6       "mode": "row-based",
7       "facets": [
```

```

8      {
9        "selectNumeric": true ,
10       "expression": "grel:diff(cells['DateOfClaim'].value, cells['
11         PolicyStartDate'].value, \"days\")",
12       "selectBlank": true ,
13       "selectNonNumeric": true ,
14       "selectError": true ,
15       "name": "DateOfClaim" ,
16       "from": -1000000,
17       "to": 0,
18       "type": "range" ,
19       "columnName": "DateOfClaim"
20     }
21   ]
22 }
23 ]

```

-
- A check to see that the PolicyStartDate is in the past. The OpenRefine API call for this is shown in Appendix H.
 - A check to see that the ExcessPaid, ClaimAmount and AmountPaid is greater than zero. An example of this OpenRefine API call for ExcessPaid is shown in Appendix I. This was repeated for ClaimAmount and AmountPaid.

After the data quality check had been enforced, outliers were removed from the data set. The outliers were determined by calculating the Z scores of each continuous variable. For logistic regression, the data was filtered to remove the DateDifference, AmountPaid, Age, TotalPoliciesRevenue, SumInsured and ExcessPaid outliers and an example of the R code is shown below for DateDifference, where claimsIn is the claims data set.

```

1 claimsIn$DateDiffZ <-scale(claimsIn$DateDiff)
2 claimsIn =subset(claimsIn , DateDiffZ >=-3.29)
3 claimsIn =subset(claimsIn , DateDiffZ <=3.29)

```

8.1.4 Filtering the Test Data

8.1.4.1 Filtering of Fields for Logistic Regression

Because logistic regression works best with continuous data, the categorical variables such as Broker and Insurer were removed from the training data set. Logistic regression can be overfitted and hence, the most important continuous variables were used to train the ICFPM model. This resulted in FraudulentClaimIndicator, DateDifference, AmountPaid, Age, TotalPoliciesRevenue, SumInsured and ExcessPaid. Other continuous variables could have been used but these were determined to work well with this research.

8.1.4.2 Filtering of Fields for Apriori Association Rules

The fields that were available for the original data set were used to generate the Apriori association rules. Continuous data was not filtered out as one would expect, instead, transformation was done – as is explained below.

8.1.5 Transforming the Test Data

The main transformation of the test data was performed for Apriori association rules. The continuous variables were transformed into discrete variables and an example of this is shown below. The example shows the PostalCode field changed into a discrete variable. The PostalCode variable is shown as a range instead of an integer. This was repeated for all continuous variables such as SumInsured, TotalPoliciesRevenue and Age.

```
1 claimsIn$PostalCode<-as.factor(gsub(",",";",cut(claimsIn$PostalCode,
breaks=c(0,1,2899,4730,6499,8299,9999))))
```

This was the main transformation performed as the filtering of fields and addition of other calculated fields do not count as transformation.

8.1.6 Generation of the ICFPM Model

From the original framework that was specified for this research, it was stipulated that the Insurance Claims Fraud Prediction Model as shown in Figure 8.8 needs to include two sub-models: the Logistic Regression Sub-Model (IFLMLM), and the Apriori Association Rules Sub-Model (IFAMLM). As previously mentioned, logistic regression was used as the supervised machine learning algorithm and Apriori association rules were used as both a supervised and an unsupervised machine learning algorithm. The training of these two sub-models are described below.

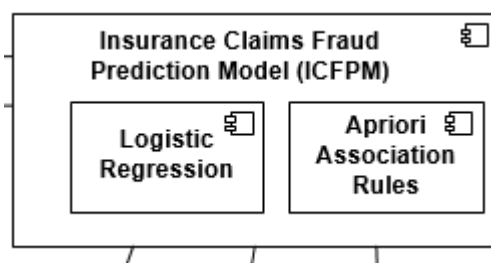


Figure 8.8: Insurance Claims Fraud Prediction Model

8.1.6.1 Training of the Insurance Claims Fraud Apriori Association Rules Machine Learning Model (IFAMLM)

To perform the tests, data was used as an input to the model to train it. To generate the unsupervised association rules, fifty thousand claims that were not indicated as fraud were added to the Apriori model (*sub-folder:5. Models\1. IFAMLM Model*). The support was set at 0.075 and confidence was set at 0.8. This generated 510 rules (*sub-folder:6. Testing\1. Rules*).

Next, the supervised association rules were generated by adding 1900 fraudulent claims and 19 000 non-fraudulent claims. The unbalanced distribution was estimated from the statistic that 10% of all property and casualty claims are fraudulent (Insurance Information Institute, 2017). As mentioned previously, the training data set needed to have at least 20 000 transactions in it (Liao and Zhu, 2014). To generate the rules, the support was set to 0.02 and the confidence to 0.75. The reason for the support being low was because the machine learning technique needed to be more sensitive to fraudulent claims. This generated 306 rules (*sub-folder:6. Testing\1. Rules*).

8.1.6.2 Training of the Insurance Claims Fraud Logistic Regression Machine Learning Model (IFLMLM)

The claims were afterwards used to train the logistic regression model with FraudulentClaimIndicator as the predicted variable and DateDifference, AmountPaid, Age, TotalPoliciesRevenue, SumInsured and Excess as the predictor variables (*sub-folder:5. Models\2. IFLMLM Model*).

8.2 Validation of the Prototype

8.2.1 Statistical Accuracy Testing

The accuracy of the Insurance Claims Fraud Prediction Score Model (ICFPSM) was then tested by using a package of R called *accuracy.meas* (Lunardon, 2016). This is because the model was created using unbalanced data and as such, this R accuracy package is valuable.

The package was used to calculate the measures of accuracy, Precision, Recall and F-measure. These measures are described as follows:

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

(Davis and Goadrich, 2006)

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives}$$

(Davis and Goadrich, 2006)

$$FMeasure = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

(Powers, 2011)

From these three measures of accuracy, it is important to note that the F-measure is the harmonic mean between recall and precision (Sasaki et al., 2007).

To calculate these measures for the proposed model, the ICFPM was tested using the fraud prediction threshold at 15%, which means that any claim with a suspicion factor of 15% or higher needed to be investigated. To test the model, a data set was created with the fields shown in Table 8.1 (*sub-folder:6. Testing\2. Data For F-test*).

Table 8.1: Table Showing Data Used for Creating the F-test

Field Number	Field Name
1	Number of Supervised Rules Fired
2	Number of Unsupervised Rules Broken
3	Probability Calculated using Logistic Regression
4	Insurance Claims Fraud Prediction Score
5	Over-Fraud-Threshold Indicator
6	Actual Fraud Indicator

Based on this table, the approach was applied that any claim with a total fraud prediction score over the suspicion factor would need to be investigated. Hence, to determine the accuracy of using this method, one would need to use the over-fraud-threshold indicator and the actual fraud indicator.

This test was done with 10 000 claims of which 1000 were fraudulent. The unbalanced distribution was once again estimated from the statistic that 10% of all property and casualty claims are fraudulent (Insurance Information Institute, 2017). These claims were specifically engineered to include generic non-fraudulent claims and claims that had been specifically added with fraudulent characteristics as have been described in this research report.

To calculate an F-measure, a table of three arguments is needed (Lunardon, 2016). Table 8.2 shows these arguments with what they would contain with regard to this research and the related insurance claims.

Table 8.2: Table Showing the Requirements of an F-test

Argument Number	Argument	Contents
1	Response	This contains a vector of the actual fraudulent claim indicators.
2	Predicted	This contains a vector with the insurance claims fraud prediction score.
3	Threshold	This is the threshold that classifies a claim as positive for being fraudulent.

From the data set created with this information, the accuracy of the ICFPM model was tested (*sub-folder:6. Testing*). It was determined that the precision of the model was equal to 0.616 and the recall was equal to 0.762.

From these two measures of accuracy, the F-measure could be calculated as follows:

$$FMeasure = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

$$FMeasure = 2 \cdot \frac{0.616 \cdot 0.762}{0.616 + 0.762}$$

$$FMeasure = 0.681$$

The high measure of precision and recall mean that the false positive rate of the ICFPM model is low and the false negative rate of this model is also low. Since both the precision and recall measures are relatively high, one can specify that the ICFPM model generated has an adequate accuracy and is valuable in the prediction of fraud.

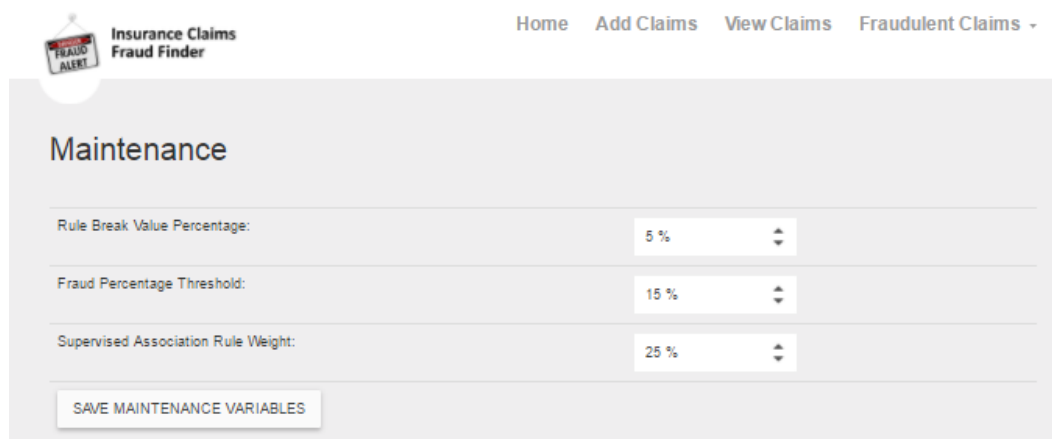
8.2.2 Scenario-based Testing

After the statistical accuracy testing, it is valuable to depict scenarios that show how claims would be predicted as fraudulent. The first scenario shows when supervised

machine learning rules increase the chance of fraud enough to justify the claim being investigated. In the second scenario, unsupervised machine learning rules also increase the chance of fraud enough for the claim to be investigated. In the last scenario, logistic regression also increases the chance of fraud enough for the claim to be investigated. It is important to note that for all three scenarios, supervised Apriori association rules, unsupervised association rules and logistic regression were applied to the claim. However, each scenario focuses on which one makes the claim flagged as a high likelihood of fraud.

8.2.2.1 Predicting Suspicious Transactions and Fraud

For the following scenarios, the maintenance variables of the system are set as shown in Figure 8.9.



The screenshot shows a web application interface for 'Insurance Claims Fraud Finder'. At the top, there is a navigation menu with 'Home', 'Add Claims', 'View Claims', and 'Fraudulent Claims'. Below the navigation, the 'Maintenance' section is displayed. It contains three rows of settings, each with a label and a dropdown menu:

- Rule Break Value Percentage: 5 %
- Fraud Percentage Threshold: 15 %
- Supervised Association Rule Weight: 25 %

At the bottom of the maintenance section, there is a button labeled 'SAVE MAINTENANCE VARIABLES'.

Figure 8.9: Maintenance Variables

As described in Table 6.1, it is evident that if the total ICFPS score of the claim is greater than 15%, the claim must be flagged as having a high chance of fraud and therefore it must be investigated.

The following scenarios show the value of using such a system to predict fraudulent insurance claims.

Scenario 1

The first scenario manifests when enough supervised machine learning rules are triggered to make an impact on the Insurance Claims Fraud Prediction Score (ICFPS). The machine learning rules are classified as supervised when the consequent of the rule is that the fraudulent claim indicator is equal to true. PPDM (Privacy Preserving Data Mining) is excluded from the following scenario so that the results are easier to read; however, PPDM results are shown later in this chapter. The account number is hashed.

1. An instance occurred where an agent from an insurance brokerage submitted fraudulent claims without the policyholders' knowledge. The agent submitted claims on behalf of policyholders who had not claimed from their insurers. Before the claim was paid, the agent changed the account details of the policyholder on the system to his own banking details. Once the claim was paid, the suspect changed these details back (SAICB, 2014).
2. Based on this information, claims would exist in the data set with an agent ("Agent Z") of an insurance brokerage ("Broker X") through "Insurer Y", with a policyholder's account number being changed to "9048EAD9080D9B27D6B2B6ED363CBF8CCE795F7F" on the claim so that the actual policyholder would not be paid. After some time, claims processed by this agent were determined to be fraudulent. There are therefore claims in the model training data set including the fields as shown in Table 8.3.
3. This resulted in the sub-set of rules shown in Table 8.4 that are specific to the aforementioned claims.
4. The same syndicate moved their operation to "Broker D" through "Insurer E", with the agent changing his name to "Agent F".
5. The agent involved with the syndicate processed a claim and although the agent's alias, broker and insurer had changed, the agent captured the same account number on the system. If the claim had passed through the claims management system that used this implementation of the ICFPSM model to intelligently predict insurance claims fraud, the claim would have been flagged as fraudulent and could have been investigated.

Table 8.3: Table Showing Field Values in Training Data-Set to Prove Scenario 1

Field	Value
Agent	Agent Z
Broker	Broker X
Insurer	Insurer Y
Payment Account Number	9048EAD9080D9B27D6B2 B6ED363CBF8CCE795F7F

Table 8.4: Table Showing Subset of Rules for Scenario 1

Rule
$\{\text{brokerid} = \text{Broker X}, \text{agentid} = \text{Agent Z}\} \Rightarrow \{\text{fraudulentclaimindicator} = 1\}$
$\{\text{paymentaccountnumber} = 9048EAD9080D9B27D6B2B6ED363CBF8CCE795F7F\} \Rightarrow \{\text{fraudulentclaimindicator} = 1\}$
$\{\text{brokerid} = \text{Broker X}, \text{agentid} = \text{Agent Z}, \text{paymentaccountnumber} = 9048EAD9080D9B27D6B2B6ED363CBF8CCE795F7F\} \Rightarrow \{\text{fraudulentclaimindicator} = 1\}$
$\{\text{agentid} = \text{Agent Z}\} \Rightarrow \{\text{fraudulentclaimindicator} = 1\}$
$\{\text{paymentaccountnumber} = 9048EAD9080D9B27D6B2B6ED363CBF8CCE795F7F\} \Rightarrow \{\text{fraudulentclaimindicator} = 1\}$

From this scenario, it is evident that generating a shared data set among insurers and brokers can add value. The addition of machine learning to this Big Data file system will result in shared knowledge amongst brokerages.

Scenario 2

The possibility of fraud also arises if there is too much wrong information about the claim which means that the claim is fundamentally incorrect – as is shown in the following example. A sample of the rules generated is shown in Table 8.5. These rules have not been hashed (as they should be for this research), in order to provide a better understanding of the scenario.

Table 8.5: Random Sample of Association Rules

Rule	Support	Confidence	Lift
{policyholdercity=Rustenburg} => {province=North West}	0.015872	1	11.546157
{area=Johannesburg} => {postalcode=(0;2.9e+03]}	0.018826	1	1.9502265
{suminsured=(2e+03;1e+04], probableamountofentireloss=(2e+03;1e+04]} => {amountclaimed=(2e+03;1e+04]}	0.023690	0.964715	20.118375
{gender=male, suminsured=(2e+03;1e+04]} => {amountclaimed=(2e+03;1e+04]}	0.017034	0.966480	20.155185
{maritalstatus=Single, suminsured=(2e+03;1e+04]} => {amountclaimed=(2e+03;1e+04]}	0.019870	0.973938	20.310711
{suminsured=(5e+05;1e+06], amountclaimed=(2e+05;5e+05]} => {probableamountofentireloss=(5e+05;1e+06]}	0.016167	1	21.149521
{suminsured=(5e+05;1e+06], amountclaimed=(2e+05;5e+05]} => {probableamountofentireloss=(5e+05;1e+06]}	0.016167	1	21.149521
{probableamountofentireloss=(1e+04;2e+04], amountclaimed=(1e+04;2e+04]} => {suminsured=(1e+04;2e+04]}	0.036845	1	11.779169
{suminsured=(4e+04;5e+04], probableamountofentireloss=(4e+04;5e+04]} => {amountclaimed=(3e+04;4e+04]}	0.022922	0.960396	9.3660285
{kindofloss=Theft - Household, amountclaimed=(4e+04;5e+04]} => {suminsured=(5e+04;1e+05]}	0.029460	0.991385	4.3541369

Noteworthy fields when inputting the claims information:

Total policies revenue:	5,050.00	⬆️⬇️⬆️
Amount claimed:	20,000.00	⬆️⬇️⬆️
Policy start date:	01/01/2009	📅
Date of claim:	31/03/2017	📅
Postal code:	0044	
City:	Durban	
Policyholder's province:	Gauteng	

Figure 8.10: Inputted Fields for the Claim with Fundamental Issues

Once this claim was added to the claims management system, it was passed through the supervised rule engine where no rules were triggered as they did not include the same information. Three unsupervised rules were however broken, as shown below:

Rules Broken
<code>{city=Durban} => {policyholderprovince=KwaZulu-Natal}</code>
<code>{city=Durban} => {postalcode=(2.0e+03;4.73e+03)}</code>
<code>{fraudulentclaimindicator=0, totalpoliciesrevenue={5e+04;1e+05}, amountclaimed={1e+04;2e+04}} => {date_diff=2.0e+03;4e+03}</code>

Figure 8.11: Unsupervised Rules Broken for the Claim with Fundamental Issues

These rules increased the ICFPS score for each broken rule. Logistic regression did not yield a substantial probability that the claim was fraudulent, but it did add 0.1% to the ICFPS score. This resulted in the claim getting outputted as possibly fraudulent and indicated that it warranted investigation.

Scenario 3

The results gained from predicting fraud by using the Insurance Claims Fraud Logistic Regression Machine Learning Model (IFLMLM) added to the total chance of fraud. Since the output of logistic regression was a probability, the score from logistic regression in this scenario was also referred to as a probability. The results of each variable in the effect plots generated from R can be noted in Figure 8.12.

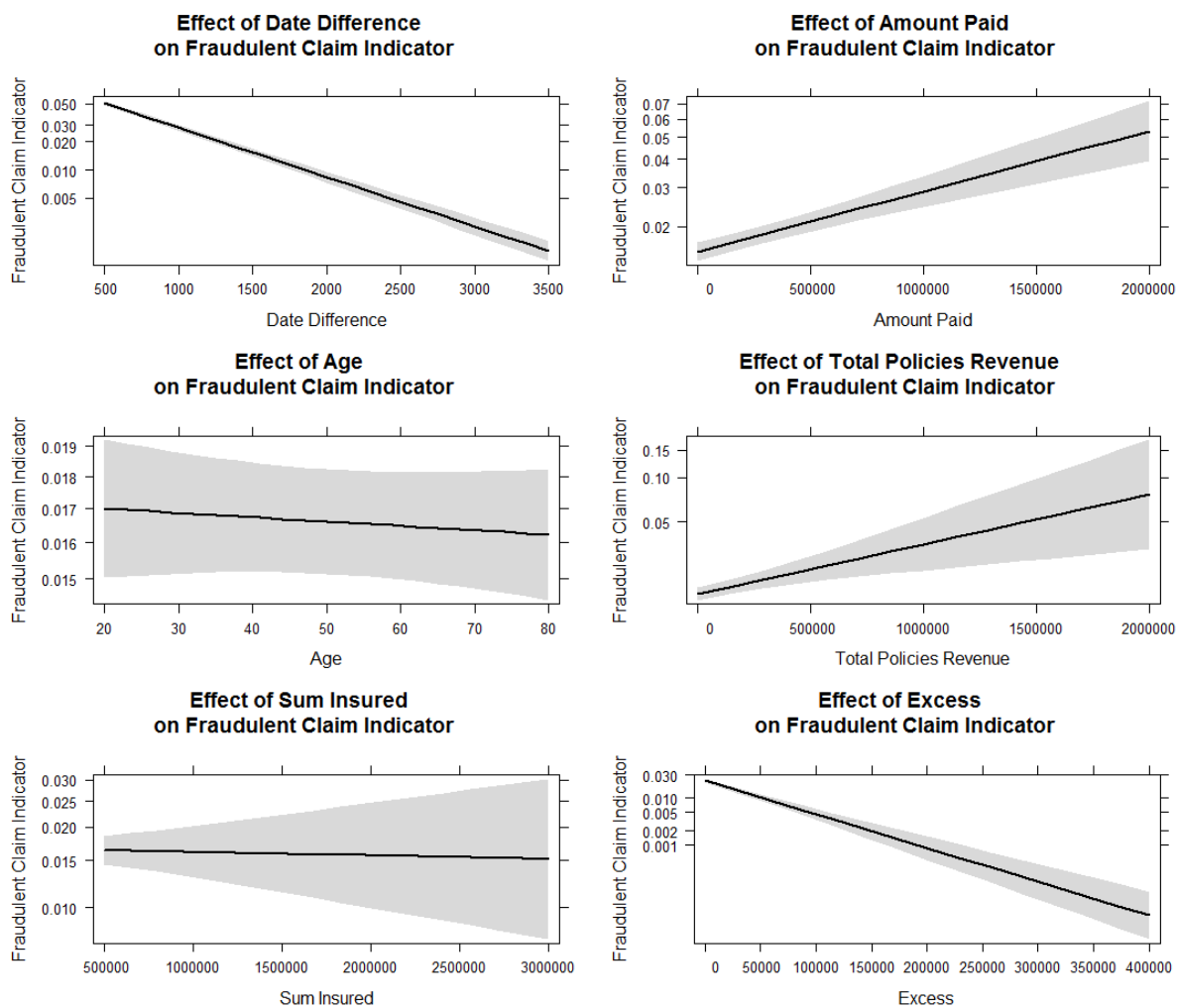


Figure 8.12: Logistic Regression Results

The scenario that was affected most by logistic regression occurred when the value predicted from logistic regression resulted in the claim being over the Fraud Percentage Threshold.

To test what possible claims would be over this threshold, the code shown in Appendix B was used. It generated 7 605 000 claims of which only the fields needed for logistic regression were included. These claims were subsequently passed through the Insurance Claims Fraud Logistic Regression Machine Learning Model (IFLMLM), and the results in Table 8.6 show a five-number summary of the results of generating these claims and passing them through the model. A five-number summary is a commonly used representation of the descriptive statistics: Min, Max, 1st Quartile, 3rd Quartile and Median (Mendenhall et al., 2012, pp. 80).

Table 8.6: Five Number Summary of Fabricated Claims Passed Through Logistic Regression

	Amount Paid	Age	Total Policies Revenue	Date Difference	Sum Insured	Excess	Fraudulent Claim Indicator Prediction
Min	R0.00	15	R0.00	0	R0.00	R0.00	0.0003%
1st Quartile	R150,000.00	35	R60,000.00	100	R150,000.00	R7,000.00	0.5973%
Median	R312,500.00	60	R150,000.00	1200	R312,500.00	R40,000.00	2.5176%
3rd Quartile	R475,000.00	85	R450,000.00	2300	R475,000.00	R75,000.00	8.3428%
Max	R4,500,000.00	105	R4,500,000.00	6000	R4,500,000.00	R209,000.00	85.172%

This can be compared to an actual claims set that was passed through the IFAMLM model, as is shown in Table 8.7.

Table 8.7: Five Number Summary of Actual Claims Passed Through Logistic Regression

	Amount Paid	Age	Total Policies Revenue	Date Difference	Sum Insured	Excess	Fraudulent Claim Indicator Prediction
Min	R0.00	19	R0.00	1	R0.00	R0.00	0.0185%
1st Quartile	R4,500.00	35	R11,790.00	386	R31,422.00	R2,800.00	0.5671%
Median	R26,600.00	51	R37,631.00	1090	R64,038.00	R16,432.00	2.6195%
3rd Quartile	R116,241.00	68	R97,595.00	2519	R157,760.00	R68,955.60	5.8884%
Max	R2,106,000.00	85	R2,502,033 .00	3716	R3,095,244.00	R1,263,600.00	26.677%

Since the actual claims are more likely to occur and do not follow an unusual pattern, they are not predicted to be fraudulent as often. The claims in the fabricated data set are flagged as fraudulent more often due to their unusual nature. It can be noticed from the results set that claims that have been paid with a large difference between the paid amount and the insured value are flagged with the highest probability of being fraudulent. The types of claims that come up as second most suspicious are claims with a low difference between the start date of the policy and the claimed date.

Once these probabilities are added to the total ICFPS score (which includes the scores for the association rules), the claim can be indicated as possibly fraudulent.

8.2.3 PPDM Scenario Test

Because the claims data set used to create the ICFPM model and predict insurance claims fraud contain personal information such as names, surnames and addresses, it is considered vital to employ PPDM to prevent a breach of this data. An example of why this is necessary is when association rules are potentially created from this personal information. It would constitute a data breach if these rules were provided to insurers and brokers with valuable information that can be used to identify people. To prove that PPDM is valuable and viable, the researcher puts forth a hypothetical scenario:

1. Insurance was applied for by a criminal syndicate at “Broker A” through “Insurer A” with the name “John”, surname “Smith” and the agent on the policy was “Agent X”. This claim was found to be fraudulent and as such was not paid.
2. The exact same syndicate applied for insurance with “Broker B” at “Insurer B”, thinking that the broker and insurer would not share information. The syndicate subsequently claimed through this insurer as well. Unfortunately, when capturing the policyholder’s information, the insured’s name was captured incorrectly as “Jon”, with his surname still as “Smith”.
3. As mentioned earlier, names are standardised during pre-processing by using field standardisation algorithms. Figure 8.13 shows results of passing incorrect first names through the Levenshtein algorithm. Errata such as incorrect spelling are found and names are merged into one name.
4. After the names have been standardised, they are hashed. If the data had not been standardised prior to hashing with SHA1, the name “John” would have become “5753a498f025464d72e088a9d5d6e872” and “Jon” would have become “eb618462d8ab174344edea26bbab6f70”. These values are vastly different due to the nature of SHA1 hashing. This would mean that an association rule would not be triggered or broken by a claim if there was one including this information.
5. When the data with the first claim in it were initially passed through the Apriori algorithm with an extremely low support, the following rule would appear: “{surname = 96bcf8c98f94b6ace4a4b716cf0e3b32,firstname = 5753a498f025464d72e088a9d5d6e872 => fraudulentclaimindicator = true}”. This would not be an optimal result, as for this rule to develop, the support would need to be much lower than normal. It is obviously possible that there could be insured policyholders with the same personal details as an individual who committed insurance claims fraud. However, owing to the fact that the support was low for this rule, the likelihood that this person committed fraud would only be marginally higher than for another policyholder, unless other rules were also broken or triggered, or if the logistic regression resulted in a higher than normal likelihood of fraud.
6. If a data scientist is intent on finding low-level, low-support rules such as this

one, or if rules such as this actually appear in the rule-set without the data scientist’s intervention, it is clear that hashing the data before passing it through the algorithm effectively protects policyholder information.

- The rules are converted to XML so that they can be applied to new claims. When the new claim from the same syndicate with the same name is inputted into the claims management system, pre-processed and then hashed with exact same algorithm, the rule will be triggered. It is essential, however, that the hashing algorithm and salt need to be standard, otherwise the rule will not be triggered.

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	2	<ul style="list-style-type: none"> John (1 rows) Jon (1 rows) 	<input type="checkbox"/>	John
2	2	<ul style="list-style-type: none"> Jane (1 rows) Jane (1 rows) 	<input type="checkbox"/>	Jane

Figure 8.13: Text Cleaning with the Levenshtein Algorithm

Floridi (2014) maintains that data is anonymous if the person to whom the data relates cannot be re-identified from the data. Since PoPIA is relatively new legislation and terms such as anonymous data have not been covered by the legislation, meeting the needs of international legislation could result in PoPIA compliance. Because a person cannot easily be re-identified from the data used and outputted by the ICFPM model, the data can be seen to be anonymous and does not violate EU data regulations or the PoPI Act.

8.2.4 Architecture Capacity Testing

It was previously mentioned that 20,000 claims would be sufficient to train the model. It is however necessary to determine whether the model could cater for all claims records that an insurance company in South Africa could be expected to have. This was tested by training the IFAMLM and IFLMLM sub-models with different numbers of claims and recording the time taken to train these sub-models. The range of the number of claims used started at 20,000 as this was the minimum number of claims required for accuracy and ranged to 425,000 as this was the highest number of insurance claims in South Africa for an insurer at the start of this research (BusinessTech, 2015). Multiple experiments were performed in increments of 20,000 to determine the effect of the number of claims on model training time. This is shown in the table in Appendix J. The table shows the number of records used to train the IFAMLM and IFLMLM sub-models and the amount of time it took to train each one. The confidence and support for the IFAMLM model had to be kept constant at 0.075 and 0.8 respectively for the test. From these tests, the scatter chart in Figure 8.14 was plotted. It is noticeable that as the number of claim records increases, so does the time taken to train both the IFAMLM and IFLMLM sub-models.

Although the results show that there is a direct correlation between data set size and training time, it is important to note that most of the time taken to train the model was for pre-processing. Running the Apriori algorithm for 500,000 records only took 2.78s and running the logistic regression algorithm only took 0.442s. To remedy this, the data could be kept as pre-processed in the Big Data file system. This would drastically increase the storage space required as the pre-processed data is different for each algorithm.

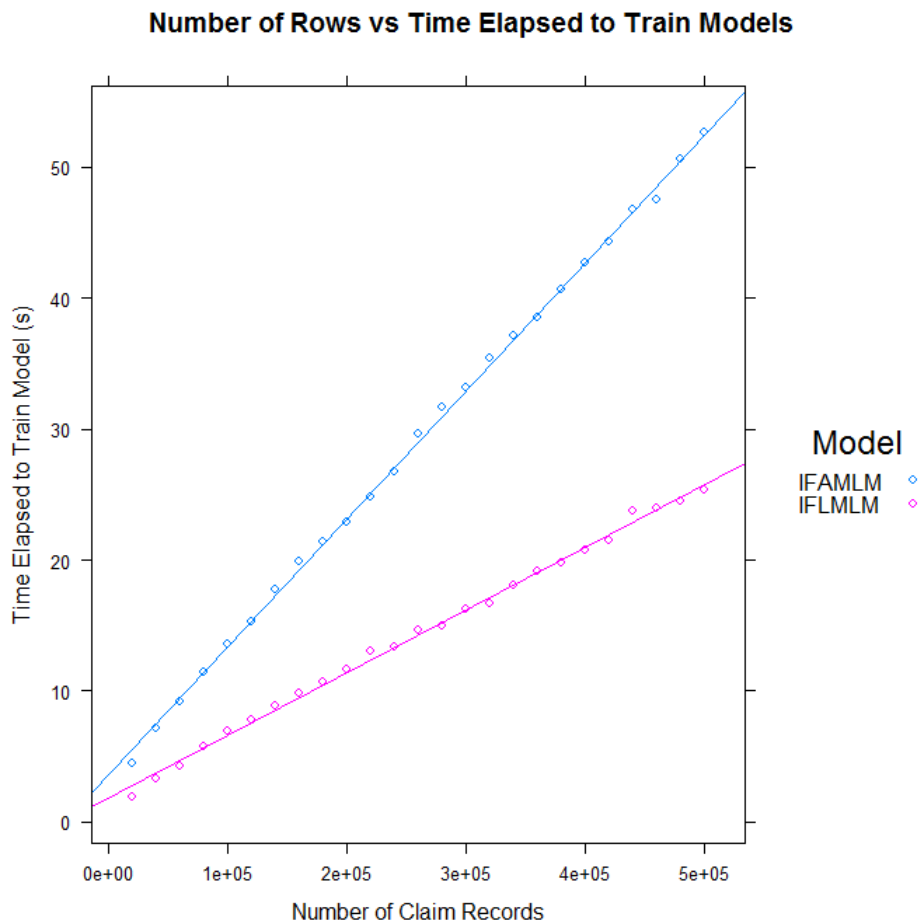


Figure 8.14: Architecture Capacity Testing: Number of Records vs Time Taken to Train the Model

8.3 Observations Gleaned regarding Fraudulent Claims

Information has been gathered from the current research regarding fraudulent claims and key factors that could possibly indicate that a claim is fraudulent. From the research presented, and based on the claims data set used, the following observations were made:

1. The sum insured of a policy has little effect on whether a claim is fraudulent or not.

2. Although the age of the policyholder has some effect on whether a claim is fraudulent or not, there is not a strong correlation between fraud and policyholder age.
3. As the difference between the start date of the policy and the date of claim increases, the chance of the claim being fraudulent decreases.
4. Low claim amounts are less likely to be fraudulent than high claim amounts. Although this factor was noted, it does not have the strongest correlation among all the rules.
5. The excess amount also has an effect on fraudulent claims. Claims that require little or no excess to be paid have a higher chance of being fraudulent.
6. Third party information did not make much of an appearance during testing and hence were removed from the training data set.

It can be beneficial for data scientists to take these observations into consideration when trying to intelligently predict insurance claims fraud. The observations above can be used to reduce the number of fields in the training data set and as a basis point to understand the kinds of correlations that should be noticed.

8.4 Discussion

This chapter showed how a prototype was built that was used to intelligently predict insurance claims fraud. It also presented tests that were performed to validate the prototype, such as statistical accuracy testing, scenario-based testing, testing to prove the efficacy of PPDM and architecture capacity testing. The tests showed that the ICFPM model that had been created was effective in intelligently predicting insurance claims fraud. It was determined during testing that information such as the third party did not aid in the prediction of insurance claims fraud and hence it was removed from the training data set. This could also have been remedied by using graph/map machine algorithms that might show correlations between policyholders and the third party. Using graph/map machine learning algorithms would be valuable for future work. The testing of PPDM showed that predicting insurance claims fraud within the bounds of privacy legislation is possible.

9 Conclusion

The research reported on in this dissertation examined how insurance claims fraud could be intelligently predicted. It provided an overview of the insurance industry, data in the insurance industry and insurance claims fraud. A proposed framework, suggested architecture and model were put forth to substantiate the use of Big Data frameworks and data science platforms to intelligently predict insurance claims fraud. This final chapter determines whether the research presented meets the criteria specified in the main research question and answers the secondary questions put forth.

9.1 Addressing the Problem Statement

The main purpose of this research was to investigate an intelligent way to predict whether insurance claims are fraudulent. This intelligent way of predicting insurance claims fraud was restricted by the fact that privacy legislation restricts how insurance data that relates to a person can be stored and what it can be used for. The proposed theorem was phrased as follows:

Intelligent methods can be used to predict insurance claims fraud. This can be performed within the bounds and constraints of privacy legislation and the fact that insurers have large sets of data.

To prove this result, there was a main question and sub-questions that needed to be answered.

9.1.1 The Main and Secondary Research Questions

The main question posed was:

What elements should a solution have so as to be used to intelligently predict insurance claims fraud?

To determine whether this question was successfully answered, it needs to be split into its component parts. These component parts can be related to the secondary research questions. Hence, it was necessary to determine what insurance claims fraud is, what

the prediction of insurance claims fraud involves, and how fraud can be intelligently predicted.

Determining what insurance claims fraud is, was explored in Chapter 2 and Chapter 3. Chapter 2 described the insurance industry and the issues that have developed in the insurance industry with regard to data. Among the many issues mentioned, insurance fraud was shown as a major problem in the insurance industry. Data in the insurance industry was also discussed, and it was showed that data can solve some of the issues in the insurance industry. Seeing that insurance fraud was found to be a large problem in the insurance industry and the focus of this research was on insurance claims fraud, Chapter 3 had a focus on fraud in the insurance industry. This was done by showing the types of fraud in the insurance industry, such as hard fraud and soft fraud, and describing examples of this fraud. This was followed by a discussion of a sub-set of insurance fraud, namely insurance claims fraud, which was identified as “the criminal deception of an insurance company to gain unjust payment relating to a loss contained in an insurance policy”. Insurance claims processes were also studied so as to foster a better understanding of when insurance claims fraud might occur.

In examining insurance claims fraud and how it can be intelligently predicted, the researcher attempted to answer the first of the secondary research questions:

Sub-question 1: What existing techniques are used to detect and predict insurance claims fraud?

To understand what the prediction of insurance claims fraud involves, a holistic view of fraud prediction in financial systems was described in Chapter 3. This was split into traditional fraud detection techniques and intelligent fraud prediction. The traditional fraud detection techniques included fraud auditing, forensic accounting and whistle-blowing. Intelligent fraud prediction was split into intelligent fraud prediction in financial systems and intelligent fraud prediction in insurance claims systems. It was shown that machine learning is commonly used in financial systems and insurance claims systems to predict fraud. It was also interesting to note that both supervised and unsupervised machine learning could be used to predict fraud, and combining more than one machine learning technique could increase prediction accuracy. It was determined that using Big Data to train data science models was effective at predicting fraud in financial systems.

Although Chapter 3 did provide a high-level understanding of using Big Data and data science technologies to predict insurance claims fraud, it revealed that further investigation was necessary to answer the next secondary research question:

Sub-question 2: Can new developments in Big Data as well as in data science help to predict insurance claims fraud?

This issue was addressed in Chapter 4, which focused on Big Data and data science. The chapter began by describing the three V’s of Big Data, namely Volume, Velocity

and Variety and proceeded to mention that Big Data requires advanced computing requirements. Commonly used Big Data frameworks include Spark, Hadoop and Dremel, but from these three, Hadoop was determined to be the best fit for this research. The chapter also described data science and mentioned the fields in data science such as machine learning, predictive analytics and data pre-processing. Since machine learning could be used to predict insurance claims fraud, the researcher evaluated both supervised and unsupervised machine learning techniques. It was determined that logistic regression and Apriori association rules could be used in this research to predict insurance claims fraud. These two machine learning techniques were also validated in Chapter 8 in the testing of the prototype. It was noted that R would be an adequate data science platform to perform machine learning for this research. Chapter 8 continued to address the third secondary research question:

Sub-question 3: Can a solution that limits how insurance data is stored and what it can be used for, be generated within the restrictions imposed by privacy legislation?

It was proposed that Privacy Preserving Data Mining (PPDM) be used to protect the privacy of policyholders when using data science to gain a result. The research mentioned anonymisation, perturbation and cryptographic methods for PPDM. It became evident that cryptographic hashing could be useful if data science techniques were to be applied to multiple insurers' and brokers' data.

The three secondary questions provided a foundation to determine “*What elements should a solution have*”, as stated in the primary research question. To answer this, a proposed framework, detailed design and architecture was developed. This proposed framework as shown in Chapter 5 was derived from the requirements of a solution that could intelligently predict insurance claims fraud. The framework included services, interfaces and rules. The services included were derived from the requirements, such as cleaning the data for machine learning methods, protecting the privacy of policyholders, and predicting insurance claims fraud through unsupervised and supervised machine learning. These services were used to create interface declarations. The rules, however, constrained the design of the solution and included information such as the fact that the training data set size should be between 20 000 and 60 000 claims. Arising from the rules, interfaces and services, a component diagram was shown which incorporated the use of machine learning algorithms such as logistic regression and Apriori association rules to predict insurance claims fraud. This was done using Big Data frameworks.

The detailed design described the processes that would be necessary when predicting insurance claims fraud. They were shown in Chapter 6 and included data preparation, model creation, knowledge application and model maintenance. The data preparation step described the use of PPDM to protect the privacy of policyholders' data, data cleaning and data extraction from an insurance claims management system. Next was the model creation phase, which described training the ICFPM model. This included the generation of Apriori association rules, the fitting of logistic regression to the data

and testing the accuracy of the ICFPM model. During knowledge application, it was stated that association rules would be converted to XML rules and applied to new claims through a rules engine in the insurance claims management system, while logistic regression would also be applied to new claims. This would give rise to automatic recommendations when new claims were added. Lastly, model maintenance required the ICFPM model to be retrained regularly to ensure better accuracy.

The architecture of this solution was described in Chapter 7. The architecture was derived by identifying the operational, technical and quality requirements of the system, which were used to derive both a software and a hardware architecture. The software architecture showed how R would be used as the data science platform of choice and how this would connect to Hadoop, which would contain the claims data set. The hardware architecture showed the logical hardware connectivity that was necessary to get the solution to function.

9.2 Main Contributions

9.2.1 Advancing the State of the Art

Addressing fraud in financial data is not new. This research aimed to bridge a gap in fraud within the financial sector with a focus on prediction instead of detection. Previous studies have yet to take into account two distinct real world restrictions simultaneously in the prediction of claims fraud in property and casualty insurance. Here we take into account both the consideration of large data sets of insurers and the recent advances in data privacy laws within South Africa and the global regulatory environment. These considerations served to broaden the scope and enhanced the contribution of this research.

Therefore, to prove that the research made a contribution to the state of the art, it is important to note the three main contributions made by the research.

The first contribution was the development of a framework that stipulates what needs to be performed to intelligently predict insurance claims fraud by specifying rules, services and interfaces. A constraint of this framework was that the method used to intelligently predict insurance claims fraud had to be conducted within the bounds of privacy legislation. Although this framework was directly applied to insurance claims fraud, it could easily be applied to other types of fraud in the financial services sector. This framework was the basis to derive the second and third contribution.

The second contribution was an architecture that was put forth regarding how to create a solution that can intelligently predict insurance claims fraud. This architecture included the hardware and software structure of the solution; namely, how the data science platform, R, can be used in conjunction with Hadoop to facilitate the prediction

of insurance claims fraud. Once again, although this architecture focused directly on insurance claims fraud, it was generic enough to use the components of the architecture and apply it to other types of fraud in the financial services sector.

The third contribution involved the definition of the ICFPM model. This model mathematically declared how to intelligently predict insurance claims fraud by using supervised and unsupervised machine learning techniques. It was specific and showed the reader what was necessary to train a model that can intelligently predict insurance claims fraud. The model combined Apriori association rules and logistic regression to create scores for new insurance claims. These scores were used to determine whether the insurance claims should be investigated for fraud or not.

9.2.2 Publications

- Kenyon D., Eloff J.H.P., Big Data Science for Predicting Insurance Claims Fraud, at 16th International Information Security South Africa (ISSA) Conference, ISBN 978-1-5386-0544-8, pp 40-47, 16 - 17 August 2017, Johannesburg, South Africa.

This conference paper described using Apriori association rules with PPDM to flag potentially fraudulent insurance claims. It was done by showing the steps that are necessary to use Big Data with data science to create a trained model.

9.3 Future Research

The research performed in this dissertation was a first attempt at intelligently predicting insurance claims fraud despite the restrictions imposed by privacy legislation. It achieved the goal of the problem statement through testing a proposed framework, model and architecture. Limitations of the research that do however need to be addressed, include the following:

- The accuracy of predicting insurance claims fraud through Big Data and data science based on the size of the data set was taken from prior research. This hypothesis could be tested using different data set sizes to determine an approximate accuracy. It could then possibly show the effect of the size of data on the efficacy of the Big Data and data science solution.
- The research could further be enhanced by adding new sources of data, such as from social media feeds, telematics devices and trends data. Data from social media feeds could be linked to claims based on who the policyholder is or based on what the social media feed contains. Although this could be useful in intelligent fraud prediction, it would introduce another privacy element that would need to be investigated. Telematics devices are regularly used by insurers and as such,

the geographical and driving style data created by these devices could be linked to claims. Lastly, trends data from external data providers could be used to supplement the current data sets. This would need to be investigated though. It seems that adding new types of data could prove to be more effective in predicting insurance claims fraud, but they would need to be investigated.

- Adding another type of machine learning algorithm that would find correlations within claims (such as the policyholder's surname being the same as the third party) could increase the sensitivity of the solution. Recent research by Jonker et al. (2017) used graph theory and PageRank to determine the reputation of a bank account. This approach could be used to determine the reputation of a policyholder, based on a graph of claims. The effectiveness of using this approach should be investigated.
- Another research task to enhance the process would require broker and insurer input over a period of time. This would involve brokers further marking claims as suspicious, which would result in enhanced insight in the future. Through supervised machine learning, the sensitivity of the ICFPM model could increase if brokers and insurers were to mark claims that are not proven to be fraudulent as suspicious. This practice would however need to be tested.
- Although this research did incorporate privacy and intended on protecting the privacy of policyholders, further research can be performed on PPDM. The current research hashed policyholders' data so that they could not be identified, but an in-depth study into the protection of policyholders' privacy could be performed. It could be ascertained whether policyholders could still be identified from the data that was not hashed, such as their total policy revenue and the amounts they claimed.

References

- Abraham, A. (2005). Rule-based expert systems. *Handbook of measuring system design*.
- Abraham, K. S. (2012). Four conceptions of insurance. *U. Pa. L. Rev.*, 161:653.
- ACORD (2015). A framework for the future. [Online] Available from: <https://www.acord.org/standards/framework/Pages/default.aspx>. [Accessed: 30/03/2017].
- Agrawal, R., Srikant, R., et al. (1994). Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference on Very Large Data Bases*, volume 1215, pages 487–499.
- Agresti, A. and Kateri, M. (2011). Categorical data analysis. In *International encyclopedia of statistical science*, pages 206–208. Springer.
- Al-Aidaros, K. M., Bakar, A. A., and Othman, Z. (2010). Naive Bayes variants in classification learning. In *Information Retrieval & Knowledge Management, (CAMP), 2010 International Conference on*, pages 276–281. IEEE.
- Al-Fedaghi, S. (2006). Aspects of personal information theory. In *Proceedings of the 2006 IEEE Workshop on Information Assurance*.
- Andrews, R. (2013). The big data rush: how data analytics can yield underwriting gold. [Online] Available from: <https://www.ordnancesurvey.co.uk/about/news/2013/the-big-data-rush.html>. [Accessed: 15/04/2017].
- Antón, A. I., Earp, J. B., and Young, J. D. (2010). How internet users' privacy concerns have evolved since 2002. *IEEE Security & Privacy*, 8(1).
- Aravinth, M. S., Shanmugapriyaa, M., Sowmya, M., and Arun, M. (2015). An efficient hadoop frameworks sqoop and ambari for big data processing. *International Journal for Innovative Research in Science and Technology*, 1(10):252–255.
- Arora, S. (2017). Analyzing mobile phone usage using clustering in spark mllib and pig. *International Journal of Advanced Research in Computer Science*, 8(1).
- Astute (2013). Stride, acord and astute: Working together for the insurance industry. [Online] Available from: <https://www.acord.org/AFSAfrica/Presentations/Biddie> [Accessed: 23/10/2017].
- Bachmann, F., Bass, L., Carriere, J., Clements, P., Garlan, D., Ivers, J., Nord, R., and Little, R. (2000). Software architecture documentation in practice: Documenting

- architectural layers. Technical report, Carnegie-Mellon Univ Pittsburgh Pa Software Engineering Inst.
- Bansal, M., Grover, D., and Sharma, D. (2017). Secure mining and sharing of financial data: Fuzzy logic and cryptography. *Indian Journal of Computer Science and Engineering*.
- Baranoff, E., Brockett, P., and Kahane, Y. (2012). *Enterprise and Individual Risk Management*. Saylor Academy.
- Bauer, A. R., Bowman, A. L., Keyser, R. J., Mcnamara, M. N., Urminski, C. L., Youngstrom, L., and Alfred, T. (2006). Apparatus for internet on-line insurance policy service. US Patent 7,124,088.
- Bay, S. D. (2000). Multivariate discretization of continuous variables for set mining. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 315–319. ACM.
- Bay, S. D. and Pazzani, M. J. (2001). Detecting group differences: Mining contrast sets. *Data Mining and Knowledge Discovery*, 5(3):213–246.
- Bede, P. (2017). From qualitative radiological cues to machine learning: Mri-based diagnosis in neurodegeneration.
- Begoli, E. and Horey, J. (2012). Design principles for effective knowledge discovery from big data. In *Software Architecture (WICSA) and European Conference on Software Architecture (ECSA), 2012 joint working IEEE/IFIP conference on*, pages 215–218. IEEE.
- Behera, T. K. and Panigrahi, S. (2017). Credit card fraud detection using a neuro-fuzzy expert system. In *Computational Intelligence in Data Mining*, pages 835–843. Springer.
- Bell, D. (2004). Uml basics: The component diagram. *IBM Global Services*.
- Bell, J. F. (1999). Tree-based methods. *Machine learning methods for ecological applications*, pages 89–105.
- Bengler, K., Dietmayer, K., Farber, B., Maurer, M., Stiller, C., and Winner, H. (2014). Three decades of driver assistance systems: Review and future perspectives. *Intelligent Transportation Systems Magazine, IEEE*, 6(4):6–22.
- Beni, G. (2004). From swarm intelligence to swarm robotics. In *International Workshop on Swarm Robotics*, pages 1–9. Springer.
- Berthold, M. R., Cebon, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., Ohl, P., Thiel, K., and Wiswedel, B. (2009). Knime-the konstanz information miner: version 2.0 and beyond. *AcM SIGKDD explorations Newsletter*, 11(1):26–31.

- Bertino, E., Fovino, I. N., and Provenza, L. P. (2005). A framework for evaluating privacy preserving data mining algorithms. *Data Mining and Knowledge Discovery*, 11(2):121–154.
- Bharal, P. and Halfon, A. (2013). Making sense of big data in insurance. [Online] Available from: [http://www.marklogic.com/resources/making-sense-of-big-data-ininsurance/resource download/whitepapers](http://www.marklogic.com/resources/making-sense-of-big-data-ininsurance/resource%20download/whitepapers). [Accessed: 22/05/2017].
- Bhasin, M. (2007). Forensic accounting: A new paradigm for niche consulting. *Chartered Accountant*.
- Bhowmik, R. (2008). Data mining techniques in fraud detection. *The Journal of Digital Forensics, Security and Law: JDFSL*, 3(2):35.
- Bhowmik, R. (2011). Detecting auto insurance fraud by data mining techniques. *Journal of Emerging Trends in Computing and Information Sciences*, 2(4):156–162.
- BMC (2016). Hadoop ecosystem and components. [Online] Available from: <http://www.bmcsoftware.co.za/guides/hadoop-ecosystem.html>. [Accessed: 04/10/2017].
- Boobier, T. (2016). *Analytics for insurance: The real business of Big Data*. John Wiley & Sons.
- Borgelt, C. and Kruse, R. (2002). Induction of association rules: Apriori implementation. In *Compstat*, pages 395–400. Springer.
- Borthakur, D. et al. (2008). HDFS architecture guide. *Hadoop Apache Project*, 53.
- Bosco, B. (2014). The keys to modernization: An insurance-focused approach. [Online] Available from: <http://www.insurancetech.com/channels/the-keys-to-modernization-an-insurance-focused-approach/a/d-id/1307020>. [Accessed: 22/05/2017].
- Bose, R. (2009). Advanced analytics: opportunities and challenges. *Industrial Management & Data Systems*, 109(2):155–172.
- Botha, J., Eloff, M., and Swart, I. (2015). Evaluation of online resources on the implementation of the protection of personal information act in South Africa. In *Iccws 2015-The Proceedings of the 10th International Conference on Cyber Warfare and Security: ICCWS2015*, page 39. Academic Conferences Limited.
- Botha, J., Grobler, M., Hahn, J., and Eloff, M. (2017). A high-level comparison between the South African protection of personal information act and international data protection laws. In *ICMLG2017 5th International Conference on Management Leadership and Governance*, page 57. Academic Conferences and publishing limited.
- Brennan, K. et al. (2009). *A Guide to the Business Analysis Body of Knowledge*. Iiba.
- "Broker Management System Support Manager" (2016). Broker management system captured claims data. Personal Communication.

- Brown, A., Johnston, S., and Kelly, K. (2002). Using service-oriented architecture and component-based development to build web service applications. *Rational Software Corporation*, 6.
- Brown, C. (2011). 15 most famous cases of insurance fraud. [Online] Available from: <http://www.ineffableisland.com/2011/07/15-most-famous-cases-of-insurance-fraud.html>. [Accessed: 23/10/2017].
- Brown, D. E., Abbasi, A., and Lau, R. Y. (2015). Predictive analytics. *IEEE Intelligent Systems*, (2):6–8.
- BusinessTech (2015). Best and worst insurance companies in South Africa. [Online] Available from: <https://businesstech.co.za/news/banking/88778/best-and-worst-insurance-companies-in-south-africa/>. [Accessed: 26/09/2017].
- CAIF (2016). 8 worst insurance criminals of 2016. [Online] Available from: <http://www.insurancefraud.org/hall-of-shame.htm>. [Accessed: 06/07/2017].
- Cambridge English Dictionary (2017a). Dictionary. [Online] Available from: <https://dictionary.cambridge.org/>. [Accessed: 06/11/2017].
- Cambridge English Dictionary (2017b). "insurance claim". [Online] Available from: <http://http://dictionary.cambridge.org/dictionary/english/insurance-claim>. [Accessed: 23/10/2017].
- Cambridge English Dictionary (2017c). "prediction". [Online] Available from: <https://dictionary.cambridge.org/dictionary/english/prediction>. [Accessed: 16/11/2017].
- Carbonell, J. G., Michalski, R. S., and Mitchell, T. M. (1983). An overview of machine learning. In *Machine learning*, pages 3–23. Springer.
- Carminati, M., Caron, R., Maggi, F., Epifani, I., and Zanero, S. (2015). Banksealer: A decision support system for online banking fraud analysis and investigation. *Computers & Security*, 53:175–186.
- Caruana, R. and Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pages 161–168. ACM.
- Catanzaro, B., Fox, A., Keutzer, K., Patterson, D., Su, B.-Y., Snir, M., Olukotun, K., Hanrahan, P., and Chafi, H. (2010). Ubiquitous parallel computing from berkeley, illinois, and stanford. *IEEE micro*, 30(2).
- Catlin, T., Hartman, R., Segev, I., and Tentis, R. (2015). The making of a digital insurer. *Mckinsey & Company*.
- Čerka, P., Grigienė, J., and Sirbikytė, G. (2015). Liability for damages caused by artificial intelligence. *Computer Law & Security Review*, 31(3):376–389.

- Chakravorty, A., Wlodarczyk, T., and Rong, C. (2013). Privacy preserving data analytics for smart homes. In *Security and Privacy Workshops (SPW), 2013 IEEE*, pages 23–27. IEEE.
- Chandarana, P. and Vijayalakshmi, M. (2014). Big data analytics frameworks. In *Circuits, Systems, Communication and Information Technology Applications (CSCITA), 2014 international conference on*, pages 430–434. IEEE.
- Chen, C. P. and Zhang, C.-Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Information Sciences*, 275:314–347.
- Chen, M., Mao, S., and Liu, Y. (2014). Big data: A survey. *Mobile Networks and Applications*, 19(2):171–209.
- Chou, S., Li, W., and Sridharan, R. (2014). Democratizing data science. *Data for Good: KDD at Bloomberg*.
- Chowdhury, M. and Sadek, A. W. (2012). Advantages and limitations of artificial intelligence. *Artificial Intelligence Applications to Critical Transportation Issues*, 6.
- Clarke, R. (1999). Introduction to dataveillance and information privacy, and definitions of terms. *Roger Clarke's Dataveillance and Information Privacy Pages*.
- Cloudera (2017). CDh components. [Online] Available from: <https://www.cloudera.com/products/open-source/apache-hadoop/key-cdh-components.html>. [Accessed: 06/06/2017].
- Cramer, J. (2002). The origins of logistic regression. *Tinbergen Institute Discussion Paper*, 2002(119/4).
- Crawford, K. (2011). Six provocations for big data. *SSM*.
- Crotty, J. and Horrocks, I. (2016). Managing legacy system costs: A case study of a meta-assessment model to identify solutions in a large financial services company. *Applied Computing and Informatics*.
- Dal Pozzolo, A., Caelen, O., Le Borgne, Y.-A., Waterschoot, S., and Bontempi, G. (2014). Learned lessons in credit card fraud detection from a practitioner perspective. *Expert systems with applications*, 41(10):4915–4928.
- d'Aquin, M. and Jay, N. (2013). Interpreting data mining results with linked data for learning analytics: motivation, case study and directions. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, pages 155–164. ACM.
- Davenport, T. H. and Patil, D. (2012). Data scientist. *Harvard business review*, 90(10):70–76.
- Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240. ACM.

- De Bruyn, M. (2014). The protection of personal information (popi) act-impact on South Africa. *The International Business & Economics Research Journal (Online)*, 13(6):1315.
- Dean, J. and Ghemawat, S. (2010). Mapreduce: a flexible data processing tool. *Communications of the ACM*, 53(1):72–77.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.
- Denoeux, T. (1995). A k-nearest neighbor classification rule based on dempster-shafer theory. *IEEE transactions on systems, man, and cybernetics*, 25(5):804–813.
- Derrig, R. A. (2002). Insurance fraud. *Journal of Risk and Insurance*, 69(3):271–287.
- Dezyre (2016a). Cloudera vs. hortonworks vs. mapr - hadoop distribution comparison. [Online] Available from: <https://www.dezyre.com/article/cloudera-vs-hortonworks-vs-mapr-hadoop-distribution-comparison-/190>. [Accessed: 18/07/2017].
- Dezyre (2016b). R hadoop – a perfect match for big data. [Online] Available from: <https://www.dezyre.com/article/r-hadoop-a-perfect-match-for-big-data/292>. [Accessed: 18/08/2017].
- Dhar, V. (2013). Data science and prediction. *Communications of the ACM*, 56(12):64–73.
- Dhyani, B. and Barthwal, A. (2014). Big data analytics using hadoop. *International Journal of Computer Applications*, 108(12).
- Dionne, G., Giuliano, F., and Picard, P. (2003). Optimal auditing for insurance fraud. Technical report, CIRPEE.
- Dreiseitl, S. and Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics*, 35(5):352–359.
- Dworkin, T. M. and Baucus, M. S. (1998). Internal vs. external whistleblowers: A comparison of whistleblowing processes. *Journal of Business Ethics*, 17(12):1281–1298.
- Eckerson, W. W. (2007). Predictive analytics. *Extending the Value of Your Data Warehousing Investment. TDWI Best Practices Report*, 1:1–36.
- EIOPA (2014). Technical specification for the preparatory phase (part i). [Online] Available from: <https://eiopa.europa.eu/Publications/Standards>. [Accessed: 20/10/2017].
- El-Sappagh, S. H. A., Hendawi, A. M. A., and El Bastawissy, A. H. (2011). A proposed model for data warehouse etl processes. *Journal of King Saud University-Computer and Information Sciences*, 23(2):91–104.

- Eling, M., Schmeiser, H., and Schmit, J. T. (2007). The solvency ii process: Overview and critical analysis. *Risk management and insurance review*, 10(1):69–85.
- Engel, K.-J. and Nagel, R. (2000). A brief history of the exponential function. *One-Parameter Semigroups for Linear Evolution Equations*, pages 497–508.
- Eriksen, M. D. and Carson, J. M. (2017). A burning question: Does arson increase when local house prices decline? *Journal of Risk and Insurance*, 84(1):7–34.
- Ernst & Young (2013). Insurance industry challenges, reforms and realignment. [Online] Available from: <http://www.ey.com/Publication/vwLUAssets>. [Accessed: 10/10/2017].
- Essert, H. and Barron, E. (2015). 10 questions about the international capital standard. [Online] Available from: <http://www.pwc.com/us/en/insurance/insurance-modernization-process.html>. [Accessed: 04/08/2017].
- European Union (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union*, L119:1–88.
- Experian (2017). Insurance. [Online] Available from: <https://www.edq.com/insurance/insurance-data-quality/>. [Accessed: 06/10/2017].
- Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1997). From data mining to knowledge discovery in databases. *AI MAGAZINE*.
- Financial Services Board (2015). Treating customers fairly. [Accessed: 16/04/2017].
- Finn, R. L., Wright, D., and Friedewald, M. (2013). Seven types of privacy. In *European data protection: coming of age*, pages 3–32. Springer.
- Fix, E. and Hodges Jr, J. L. (1951). Discriminatory analysis-nonparametric discrimination: consistency properties. Technical report, California Univ Berkeley.
- Floridi, L. (2014). Open data, data protection, and group privacy. *Philosophy & Technology*, 27(1):1–3.
- Frachot, A., Georges, P., and Roncalli, T. (2001). Loss distribution approach for operational risk.
- Frachot, A. and Roncalli, T. (2002). Mixing internal and external data for managing operational risk. *Available at SSRN 1032525*.
- Frénay, B. and Verleysen, M. (2014). Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25(5):845–869.
- Freund, Y. and Schapire, R. E. (1995). A decision-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory*, pages 23–37. Springer.

- Friedman, T. and Judah, S. (2015). Critical capabilities for data quality tools. *Gartner*.
- FSB (2014). Retail distribution review 2014. [Online] Available from: <https://www.masthead.co.za/wp-content/uploads/2015/05/FSB-Retail-Distribution-Review-2014.pdf>. [Accessed: 10/06/2017].
- Fulbright, N. R. (2014). Fraudulent insurance claims and private prosecutions. [Online] Available from: <https://www.fanews.co.za/article/legal-affairs/10/general/1120/fraudulent-insurance-claims-and-private-prosecutions/17044>. [Accessed: 26/10/2017].
- Galaez, G., de Haan, R., and Taylor, M. H. (2015). Insurance modernization: How do we get there? <http://www.pwc.com/us/en/insurance/insurance-modernization-process.html>.
- Gallant, S. I. (1993). *Neural network learning and expert systems*. MIT press.
- Gartner (2016). Big data. [Online] Available from: <http://www.gartner.com/it-glossary/big-data/>. [Accessed: 30/03/2016].
- Gawel, B. and Skalna, I. (2014). Model driven architecture and classification of business rules modelling languages. In *Advances in Business ICT*, pages 123–131. Springer.
- "General Manager: Risk and Technology" (2016). A - rated insurers claims data. Personal Communication.
- Geraghty, M. (2001). Dynamic ratemaking for insurance. US Patent App. 09/992,408.
- Ghahramani, Z. (2001). An introduction to hidden Markov models and Bayesian networks. *International journal of pattern recognition and artificial intelligence*, 15(01):9–42.
- Goel, R. K. (2014). Insurance fraud and corruption in the United States. *Applied Financial Economics*, 24(4):241–246.
- Gollapudi, S. (2016). Integrating r with apache hadoop. [Online] Available from: <https://www.r-bloggers.com/integrating-r-with-apache-hadoop/>. [Accessed: 17/07/2017].
- Gopalani, S. and Arora, R. (2015). Comparing apache spark and map reduce with performance analysis using k-means. *International Journal of Computer Applications*, 113(1).
- Grabosky, P. and Duffield, G. (2001). Red flags of fraud. *Trends & Issues in Crime and Criminal Justice*, (200):1.
- Gschwendtner, K., Lienkamp, M., and Kiss, M. (2014). Prospective analysis-method for estimating the effect of advanced driver assistance systems on property damage. In *Intelligent Transportation Systems (ITSC), 2014 IEEE 17th International Conference on*, pages 372–377. IEEE.

- Gualtieri, M. and Curran, R. (2016). The forrester wave: Big data streaming analytics, q1 2016. *Forrester.com, Cambridge MA*.
- Gualtieri, M., Yuhanna, N., Kisker, H., and Murphy, D. (2014). The forrester wave: Big data hadoop solutions, q1 2014.
- Gupta, N. (2013). Artificial neural network. *Network and Complex Systems*, 3(1):24–28.
- Halverson, M. and Malhotra, R. (2015). Beyond insurance: Embracing innovation to monetize disruption. *Accenture Strategy*.
- Ham, K. (2013). Openrefine (version 2.5). <http://openrefine.org>. free, open-source tool for cleaning and transforming data. *Journal of the Medical Library Association: JMLA*, 101(3):233.
- Hardin, J., Hoerl, R., Horton, N. J., Nolan, D., Baumer, B., Hall-Holt, O., Murrell, P., Peng, R., Roback, P., Temple Lang, D., et al. (2015). Data science in statistics curricula: Preparing students to think with data. *The American Statistician*, 69(4):343–353.
- Hassan, A. K. I. and Abraham, A. (2016). Modeling insurance fraud detection using imbalanced data classification. In *Advances in Nature and Biologically Inspired Computing*, pages 117–127. Springer.
- Hbib, L. and Barka, H. (2016). Big data: Framework and issues. In *Electrical and Information Technologies (ICEIT), 2016 International Conference on*, pages 485–490. IEEE.
- Henzen, L., Carbognani, F., Aumassony, J.-P., O’Neilz, S., and Fichtner, W. (2009). Vlsi implementations of the cryptographic hash functions md6 and irrupt. In *Circuits and Systems, 2009. ISCAS 2009. IEEE International Symposium on*, pages 2914–2917. IEEE.
- Ho, S. (2014). Insurance agent fraud. *Fraud Magazine*, 2014.
- Hollard (2017). About us. [Online] Available from: <https://www.hollard.co.za/company-overview/about-us>. [Accessed: 22/09/2017].
- Hortonworks (2017). Hortonworks data platform. [Online] Available from: <https://hortonworks.com/products/data-center/hdp/>. [Accessed: 19/07/2017].
- Huai, Y., Chauhan, A., Gates, A., Hagleitner, G., Hanson, E. N., O’Malley, O., Pandey, J., Yuan, Y., Lee, R., and Zhang, X. (2014). Major technical advancements in apache hive. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pages 1235–1246. ACM.
- Hussain, K. and Prieto, E. (2016). Big data in the finance and insurance sectors. In *New Horizons for a Data-Driven Economy*, pages 209–223. Springer.
- IBM (2010). Insurance customer retention and growth. *IBM Software Group*.

- IIHS (2016). Ratings. [Online] Available from: <http://www.iihs.org/iihs/ratings>. [Accessed: 10/05/2017].
- ILASANews (2014). The insurance loss adjuster a value adding professional at a cost. [Online] Available from: <http://www.ilasa.org.za/2014/04/the-insurance-loss-adjuster-a-value-adding-professional-at-a-cost/>. [Accessed: 10/03/2017].
- Imandoust, S. B. and Bolandraftar, M. (2013). Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background. *International Journal of Engineering Research and Applications*, 3(5):605–610.
- Informatica (2015). Gartner critical capabilities for data quality tools report: Informatica scored among top four vendors in all six use-case categories. [Online] Available from: <https://www.informatica.com/gartner-critical-capabilities-data-quality.htmlfbid=PLXyPpYzmBh>. [Accessed: 07/03/2017].
- Insurance Fraud Working Group (2011). Application paper on deterring, preventing, detecting, reporting and remedying fraud in insurance. *International Association of Insurance Supervisors*.
- Insurance Information Institute (2017). Insurance fraud. [Online] Available from: <http://www.iii.org/issue-update/insurance-fraud>. [Accessed: 19/08/2017].
- Intel IT Center (2012). Peer research: Big data analytics. Technical report, Intel.
- Investopedia (2017). "insurance claim". [Online] Available from: <http://www.investopedia.com/terms/i/insuranceclaim.asp>. [Accessed: 23/10/2017].
- Islam, M., Huang, A. K., Battisha, M., Chiang, M., Srinivasan, S., Peters, C., Neumann, A., and Abdelnur, A. (2012). Oozie: towards a scalable workflow management system for hadoop. In *Proceedings of the 1st ACM SIGMOD Workshop on Scalable Workflow Execution Engines and Technologies*, page 4. ACM.
- Islam, M. K. and Srinivasan, A. (2015). *Apache Oozie: The Workflow Scheduler for Hadoop*. O'Reilly Media, Inc.
- Islam, S. and Ahmed, M. (2013). Implementation of image segmentation for natural images using clustering methods. *Int J Emerg Technol Adv Eng*, 3(3):175–80.
- Jacobs, A. (2009). The pathologies of big data. *Communications of the ACM*, 52(8):36–44.
- Jaiswal, J. K., Samikannu, R., and Paramasivam, I. (2016). Anonymization in ppdm based on data distributions and attribute relations. *Indian Journal of Science and Technology*, 9(37).
- Jauhar, J. B., Ghani, A. B. A., and Islam, R. (2016). Results and data analysis. In *Brain Drain*, pages 127–158. Springer.
- Jensen, R. R. and Shumway, J. M. (2010). Sampling out world. *Research Methods in Geography: A Critical Introduction*, 6:77.

- Johne, A. (1993). Insurance product development: managing the changes. *International Journal of Bank Marketing*, 11(3):5–14.
- Jones, M. T. (2011). Spark, an alternative for fast data analytics. *IBM Developer Works*.
- Jonker, C., Habeck, T., Park, Y., Jordens, F., and van Schaik, R. (2017). Graph analytics for real-time scoring of cross-channel transactional fraud. In *Financial Cryptography and Data Security: 20th International Conference, FC 2016, Christ Church, Barbados, February 22–26, 2016, Revised Selected Papers*, volume 9603, page 22. Springer.
- Jordan, M. I. and Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260.
- Jordon, S. (2016). Insurance fraud: 'its all over the place,' and you should care about it, officials say. *Omaha World-Herald*.
- Josephson (2009). A study of values and behavior concerning integrity: The impact of age, cynicism and high school character. Technical report, Josephson Institute of Ethics.
- Jovic, A., Brkic, K., and Bogunovic, N. (2014). An overview of free software tools for general data mining. In *Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014 37th International Convention on*, pages 1112–1117. IEEE.
- Kalaivani, R. and Chidambaram, S. (2014). Additive gaussian noise based data perturbation in multi-level trust privacy preserving data mining. *International Journal of Data Mining & Knowledge Management Process*, 4(3):21.
- Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., and Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. *IEEE transactions on pattern analysis and machine intelligence*, 24(7):881–892.
- Kaufman, L. M. (2009). Data security in the world of cloud computing. *IEEE Security & Privacy*, 7(4).
- KDnuggets, S. P. (2015). Analytics, data mining, data science software/tools used in the past 12 months. [Online] Available from: <http://www.kdnuggets.com/polls/2015/analytics-data-mining-data-science-software-used.html>. [Accessed: 20/02/2017].
- Khan, A. U. S., Akhtar, N., and Qureshi, M. N. (2014). Real-time credit-card fraud detection using artificial neural network tuned by simulated annealing algorithm. In *Proceedings of International Conference on Recent Trends in Information, Telecommunication and Computing, ITC*, pages 113–21.
- Kirlidog, M. and Asuk, C. (2012). A fraud detection approach with data mining in health insurance. *Procedia-Social and Behavioral Sciences*, 62:989–994.

- Kleinbaum, D. G. and Klein, M. (2010). *Introduction to Logistic Regression*, chapter 1, pages 1–39. Springer New York, New York, NY.
- Knoesen (2017). Concerns and expectations for 2017. [Online] Available from: <https://www.fanews.co.za/article/intermediaries-brokers/7/general/1227/concerns-and-expectations-for-2017/22383>. [Accessed: 24/10/2017].
- Kobryn, C. (2000). Modeling components and frameworks with uml. *Communications of the ACM*, 43(10):31–38.
- Kohonen, T. (1998). The self-organizing map. *Neurocomputing*, 21(1):1–6.
- Kolde, T. and Walker, K. (2015). 2015 Cas ratemaking and product management seminar. In *Survey of External Data Possibilities for Commercial Insurance*.
- Koorn, R., Bholasing, J., Pipes, S., Rotman, D., Kypreos, C., Cumming, S., van Kerckhoven, A., Hijikata, K., and Manchu, T. (2015). Big data analytics & privacy: How to resolve this paradox.
- Kopper, C. (2009). A software framework for km3net. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 602(1):107–110.
- Kou, Y., Lu, C.-T., Sirwongwattana, S., and Huang, Y.-P. (2004). Survey of fraud detection techniques. In *Networking, Sensing and Control, 2004 IEEE International Conference on*, volume 2, pages 749–754Vol.2.
- Kranacher, M.-J., Riley, R., and Wells, J. T. (2010). *Forensic accounting and fraud examination*. John Wiley & Sons.
- Krawczyk, M. (2009). The role of repetition and observability in deterring insurance fraud. *The Geneva Risk and Insurance Review*, 34(1):74–87.
- Krishnan, S., Haas, D., Franklin, M. J., and Wu, E. (2016). Towards reliable interactive data cleaning: A user survey and recommendations. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*, page 9. ACM.
- Krstajčić, D., Cvetković, R., and Majstorović, M. (2014). Towards the alignment of business and it in insurance company. *International Journal of Scientific and Research Publications*, 4(3):1–7.
- Kurgan, L. A. and Musilek, P. (2006). A survey of knowledge discovery and data mining process models. *The Knowledge Engineering Review*, 21(01):1–24.
- Kuys, P. and Zehnwirth, B. (1997). Abstracts and reviews. *Insurance: Mathematics & Economics*, 20(3):255.
- Laffey, T. (2004). Insurance fraud: cause and effect. [Online] Available from: <https://www.joc.com/insurance-fraud-cause-and-effect20040125.html>. [Accessed: 29/08/2017].

- Lakshminarayan, K., Harp, S. A., Goldman, R. P., Samad, T., et al. (1996). Imputation of missing data using machine learning techniques. In *KDD*, pages 140–145.
- Laskar, D., Lachit, G., et al. (2014). A review on privacy preservation data mining (ppdm). *International Journal of Computer Applications Technology and Research*, 3(7):403–408.
- Lee, R., Mark, K. P., and Chiu, D. K. (2007). Enhancing workflow automation in insurance underwriting processes with web services and alerts. In *System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on*, pages 64–64. IEEE.
- Legal Dictionary (2016). Insurance fraud. [Online] Available from: <http://legaldictionary.net/insurance-fraud/>. [Accessed: 30/03/2016].
- Len, B., Paul, C., and Rick, K. (2003). *Software architecture in practice*. Addison-Wesley.
- Levada, A. L., Mascarenhas, N. D., and Tannús, A. (2011). On combining higher-order map-mrf based classifiers for image labeling. *Integrated Computing Technology*, pages 25–39.
- Levine, D. M. and Stephan, D. F. (2009). *Even you can learn statistics: A guide for everyone who has ever been afraid of statistics*. FT Press.
- Li, H., He, H., and Wen, Y. (2015). Dynamic particle swarm optimization and k-means clustering algorithm for image segmentation. *Optik-International Journal for Light and Electron Optics*, 126(24):4817–4822.
- Li, J., Huang, K.-Y., Jin, J., and Shi, J. (2008). A survey on statistical methods for health care fraud detection. *Health care management science*, 11(3):275–287.
- Li, R. (2015). Top 10 data mining algorithms, explained. [Online] Available from: <http://www.kdnuggets.com/2015/05/top-10-data-mining-algorithms-explained.html>. [Accessed: 20/02/2017].
- Li, Y., Yan, C., Liu, W., and Li, M. (2016). Research and application of random forest model in mining automobile insurance fraud. In *Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), 2016 12th International Conference on*, pages 1756–1761. IEEE.
- Liao, Z. and Zhu, Y. (2014). When a classifier meets more data. *Procedia Computer Science*, 30(Supplement C):50 – 59. 1st International Conference on Data Science, ICDS 2014.
- Lin, X. (2014). Mr-apriori: Association rules algorithm based on mapreduce. In *Software Engineering and Service Science (ICSESS), 2014 5th IEEE International Conference on*, pages 141–144. IEEE.

- Lingaraju, P., Yellaswamy, K., and Sivaiah, B. (2013). Efficient data mining algorithms for mining frequent/closed/maximal itemsets. *International Journal*, 3(9).
- Liu, J.-L. and Chen, C.-L. (2012). Application of evolutionary data mining algorithms to insurance fraud prediction. In *4th. International Conference on Machine Learning and Computing*, pages 22–17.
- Liu, K., Giannella, C., and Kargupta, H. (2006). An attackers view of distance preserving maps for privacy preserving data mining. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 297–308. Springer.
- Lloyd, S. (1982). Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137.
- Lloyd, S., Mohseni, M., and Rebertrost, P. (2013). Quantum algorithms for supervised and unsupervised machine learning. *arXiv preprint arXiv:1307.0411*.
- Lloyds (2017). ACORd standards. [Online] Available from: <https://www.lloyds.com/the-market/operating-at-lloyds/exchange/acord-standards>. [Accessed: 27/09/2017].
- Longman Dictionary of Contemporary English (1995). framework. [Online] Available from: <https://www.ldoceonline.com/dictionary/framework>. [Accessed: 27/01/2018].
- Loukides, M. (2011). *What is data science?* O'Reilly Media, Inc.
- Lu, X. and Zheng, B. (2013). Distributed computing and hadoop in statistics. In *Proceedings 59th ISI World Statistics Congress*.
- Luck, R. (2014). POPI-is South Africa keeping up with international trends?
- Lunardon, N. (2016). Metrics to evaluate a classifier accuracy in imbalanced learning. [Online] Available from: <https://www.rdocumentation.org/packages/ROSE/versions/0.0-3/topics/accuracy.meas>. [Accessed: 18/08/2017].
- Maartens, J. (2017a). About deployr. [Online] Available from: <https://docs.microsoft.com/en-us/machine-learning-server/deployr/deployr-about>. [Accessed: 24/01/2018].
- Maartens, J. (2017b). About deployr's repository manager. [Online] Available from: <https://docs.microsoft.com/en-us/machine-learning-server/deployr/deployr-repository-manager-about>. [Accessed: 20/11/2017].
- Maartens, J. (2017c). Installing & configuring deployr 8.0.0. [Online] Available from: <https://docs.microsoft.com/en-us/machine-learning-server/deployr/deployr-installing-configuring>. [Accessed: 27/11/2017].
- Maartens, J. (2017d). Overview of deployr security. [Online] Available from: <https://docs.microsoft.com/en-us/machine-learning-server/deployr/deployr-security>. [Accessed: 21/02/2017].

- Maas, P., Graf, A., Bieck, C., Mäder, P., Hürlimann, M., and Baselgia, C. (2014). Big data and advanced analytics in the commercial insurance industry. *I. VW Management-Information*, 36:17–22.
- Macdonald, J. (2017). 8 extreme cases of insurance fraud. [Online] Available from: <http://www.bankrate.com/finance/insurance/8-extreme-cases-of-insurance-fraud-1.aspx>. [Accessed: 06/07/2017].
- MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.
- Mahlow, N., Wagner, J., Buchanan, B., and Buchanan, B. (2016). Process landscape and efficiency in non-life insurance claims management: an industry benchmark. *The Journal of Risk Finance*, 17(2).
- Makhafola, D. (2015). Short-term insurance challenges. [Online] Available from: <https://www.cover.co.za/the-challenges-that-short-term-insurance-has-to-overcome/>. [Accessed: 24/10/2017].
- Manes, A. (1945). Insurance crimes. *Journal of Criminal Law and Criminology (1931-1951)*, 35(1):34–42.
- MapR (2017). Keeping it safe: Security features. [Online] Available from: <http://doc.mapr.com/display/MapR/MapR+Overview>. [Accessed: 21/02/2017].
- MapR (2017). Mapr control system. [Online] Available from: <http://doc.mapr.com/display/MapR/MapR+Control+System>. [Accessed: 27/09/2017].
- MapR (2017). Mapr converged data platform. [Online] Available from: <https://mapr.com/products/>. [Accessed: 19/07/2017].
- MapR (2017). Preparing each node. [Online] Available from: <http://doc.mapr.com/display/MapR/Preparing+Each+Node>. [Accessed: 27/09/2017].
- Marriott, F. H. C. (2013). *Basic mathematics for the biological and social sciences*. Elsevier.
- Martin, M. and Hayes, M. (2013). Operational risk management : practical implications for the South African insurance industry. *South African Actuarial Journal*.
- Masters, M. (2009). An introduction to the business analysis body of knowledge (babok 2.0). [Online] Available from: <http://www.modernanalyst.com/Resources/Articles/tabid/115/ID/1187/An-Introduction-to-the-Business-Analysis-Body-of-Knowledge-BABOK-20.aspx>. [Accessed: 03/01/2018].

- Mazur, J. (2011). Anatomy of the ACORd TXlife XML standard. [Online] Available from: <http://www.ibm.com/developerworks/library/x-ind-acordtxlife/>. [Accessed: 06/05/2017].
- McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133.
- McGraw, G. (2004). Software security. *IEEE Security & Privacy*, 2(2):80–83.
- McGuire, M. and Dowling, S. (2013). Cyber crime: A review of the evidence. *Summary of key findings and implications. Home Office Research report*, 75.
- McNicholas, P. D., Murphy, T. B., and ORegan, M. (2008). Standardising the lift of an association rule. *Computational Statistics & Data Analysis*, 52(10):4712–4721.
- Meier, J., Hill, D., Homer, A., Taylor, J., Bansode, P., Wall, L., Boucher Jr., R., and Bogawat, A. A. (2009). What is software architecture? *Microsoft Application Architecture Guide, 2nd Edition*.
- Melnik, S., Gubarev, A., Long, J. J., Romer, G., Shivakumar, S., Tolton, M., and Vassilakis, T. (2010). Dremel: Interactive analysis of web-scale datasets. In *Proc. of the 36th Int'l Conf on Very Large Data Bases*, pages 330–339.
- Mendenhall, W., Beaver, R. J., and Beaver, B. M. (2012). *Introduction to probability and statistics*. Cengage Learning.
- Mercer (2014). Business and workforce challenges in the global insurance industry. [Online] Available from: <https://www.mercer.com/content/dam/mercer/attachments/global/mercer-business-and-workforce-challenges-in-the-global-insurance-industry.pdf>. Accessed On: 10/10/2017.
- Meyer, D. and Wien, F. T. (2015). Support vector machines. *The Interface to libsvm in package e1071*.
- Miles, M. B. and Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook*. sage. pp 18.
- Mills, S. and Forder, S. (2012). Big data brings big opportunities for insurers. *IBM Smarter Analytics*.
- Minanovic, A., Gabelica, H., and Krstic, Z. (2014). Big data and sentiment analysis using knime: Online reviews vs. social media. In *Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014 37th International Convention on*, pages 1464–1468. IEEE.
- Minnaar, J. (2000). Understanding fraud and white collar crime - the origin, definition and elements of fraud. [Online] Available from: <https://www.tei.org.za/index.php/resources/articles/fraud-and-corruption/1712-understanding-fraud-and-white-collar-crime-the-origin-definition-and-elements-of-fraud>. [Accessed: 26/10/2017].

- Mohn, R. (2013). Forensic accounting in catastrophic business interruption claims. [Online] Available from: <http://www.propertycasualty360.com/2013/04/19/forensic-accounting-in-catastrophic-business-inter>. [Accessed: 06/06/2017].
- Moon, T. K. (1996). The expectation-maximization algorithm. *IEEE Signal processing magazine*, 13(6):47–60.
- Moore, M. and Zubry, D. (2013). Collision avoidance features: initial results. In *23rd International Conference on the Enhanced Safety of Vehicles*.
- Morgenstern, A., Antonino, P., Kuhn, T., Pschorn, P., and Kallweit, B. (2017). Modeling embedded systems using a tailored view framework and architecture modeling constraints. In *Proceedings of the 11th European Conference on Software Architecture: Companion Proceedings*, pages 180–186. ACM.
- Morton, S. (1993). Strategic auditing for fraud. *Accounting Review*, pages 825–839.
- Nack, W. (1992). Blood money. *Sports Illustrated*.
- Nath, I. (2016). Fight insurance fraud by sharing data through blockchain. *IBM*.
- Nelles, O. (2013). *Nonlinear system identification: from classical approaches to neural networks and fuzzy models*. Springer Science & Business Media.
- Ngai, E., Hu, Y., Wong, Y., Chen, Y., and Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, 50(3):559–569.
- Nian, K., Zhang, H., Tayal, A., Coleman, T., and Li, Y. (2016). Auto insurance fraud detection using unsupervised spectral ranking for anomaly. *The Journal of Finance and Data Science*, 2(1):58–75.
- Nilsen, P. (2015). Making sense of implementation theories, models and frameworks. *Implementation Science*, 10(1):53.
- Noyes, K. (2015). Five things you need to know about hadoop v. apache spark. *Infor-World*.
- Nunns, J. (2016). Spark, hydra & bigquery: 5 enterprise alternatives to hadoop. [Online] Available from: <https://www.cbronline.com/cloud/spark-hydra-bigquery-5-enterprise-alternatives-to-hadoop-4891737/>. [Accessed: 19/08/2017].
- Nyce, C. and CPCU, A. (2007). Predictive analytics white paper. *American Institute for CPCU. Insurance Institute of America*, pages 9–10.
- Olalekan Yusuf, T. and Ajemunigbohun, S. (2015). Effectiveness, efficiency and promptness of claims handling process within the nigerian insurance industry. 10.
- Oracle (2016). Improving insurer performance with big data. *Enterprise Architecture White Paper*.

- Ordonez, C. (2011). Data set preprocessing and transformation in a database system. *Intelligent Data Analysis*, 15(4):613–631.
- Ormerod, T., Morley, N., Ball, L., Langley, C., and Spenser, C. (2003). Using ethnography to design a mass detection tool (mdt) for the early discovery of insurance fraud. In *CHI'03 Extended Abstracts on Human Factors in Computing Systems*, pages 650–651. ACM.
- Orriols-Puig, A., Casillas, J., and Bernadó-Mansilla, E. (2008). A comparative study of several genetic-based supervised learning systems. *Learning Classifier Systems in Data Mining*, pages 205–230.
- Othman, R., Aris, N. A., Mardziah, A., Zainan, N., and Amin, N. M. (2015). Fraud detection and prevention methods in the malaysian public sector: Accountants and internal auditors perceptions. *Procedia Economics and Finance*, 28:59–67.
- Oxford English Dictionary Online (2017a). "architecture, adj.". [Online] Available from: <http://www.oed.com.uplib.idm.oclc.org/view/Entry/10408?rskey=2V8IDA&result=1&isAdvanced=falseid>. [Accessed: 27/09/2017].
- Oxford English Dictionary Online (2017b). "architecure, n.". [Online] Available from: <http://www.oed.com.uplib.idm.oclc.org/view/Entry/10408?rskey=vY4m57&result=1&isAdvanced=falseid>. [Accessed: 14/07/2017].
- Oxford English Dictionary Online (2017c). "detect, v.". [Online] Available from: <http://www.oed.com.uplib.idm.oclc.org/view/Entry/51191?rskey=gTrh7G&result=2&isAdvanced=falseid>. [Accessed: 12/07/2017].
- Oxford English Dictionary Online (2017d). "fraud, n.". [Online] Available from: <http://0-www.oed.com.innopac.up.ac.za/view/Entry/74298?rskey=ec0aDN&result=1>. [Accessed: 10/02/2017].
- Oxford English Dictionary Online (2017e). "fraud, n.". [Online] Available from: <http://www.oed.com/view/Entry/74298?rskey=MjPN1E&result=1&isAdvanced=false>. [Accessed: 06/03/2017].
- Oxford English Dictionary Online (2017f). "intelligent, adj.". [Online] Available from: <http://www.oed.com.uplib.idm.oclc.org/view/Entry/97402redirectedFrom=intelligent>. [Accessed: 12/07/2017].
- Oxford English Dictionary Online (2017g). "predict, v.". [Online] Available from: <http://www.oed.com.uplib.idm.oclc.org/view/Entry/149856?rskey=vbCBPl&result=3eid>. [Accessed: 12/07/2017].
- Oxford English Dictionary Online (2017h). "prediction, n.". [Online] Available from: <http://www.oed.com.uplib.idm.oclc.org/view/Entry/149860rskey=MWPajLresult=1&isAdvanced=falseid>. [Accessed: 12/07/2017].

- Oxford English Dictionary Online (2017i). "regulation, n.". [Online] Available from: <http://www.oed.com.uplib.idm.oclc.org/view/Entry/161427?redirectedFrom=regulationeid>. [Accessed: 27/09/2017].
- Oxford English Dictionary Online (2017j). "traditional, adj.". [Online] Available from: <http://http://www.oed.com.uplib.idm.oclc.org/view/Entry/204304?redirectedFrom=traditional eid>. [Accessed: 27/09/2017].
- Oxford English Dictionary Online (2017k). Welcome. [Online] Available from: <http://www.oed.com.uplib.idm.oclc.org/>. [Accessed: 06/11/2017].
- Oxford English Dictionary Online (2018). "model, n. and adj.". [Online] Available from: <http://www.oed.com/view/Entry/120577?rskey=vwhSg0&result=1&isAdvanced=false eid>. [Accessed: 06/03/2018].
- Panda, M. and Patra, M. R. (2007). Network intrusion detection using naive bayes. *International journal of computer science and network security*, 7(12):258–263.
- Parmar, K. and Shah, V. (2016). A review on data anonymization in privacy preserving data mining. *International Journal Of Advanced Research In Computer And Communication Engineering*, 5(2).
- Pattalwar, S. and Agrawal, P. (2017). Approaches of privacy preservation in data mining. *International Journal*, 2(5).
- Pavlo, A., Paulson, E., Rasin, A., Abadi, D. J., DeWitt, D. J., Madden, S., and Stonebraker, M. (2009). A comparison of approaches to large-scale data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, pages 165–178. ACM.
- Peng, C.-Y. J., Lee, K. L., and Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The journal of educational research*, 96(1):3–14.
- Peterson, L. E. (2009). K-nearest neighbor. *Scholarpedia*, 4(2):1883.
- Phan, D., Beugnard, A., et al. (2001). Moduleco, a multi-agent modular framework, for the simulation of network effects and population dynamics in social sciences, market & organisations. *Approches Connexionnistes en Sciences Economiques et de Gestion*, 8.
- Phua, C., Lee, V., Smith, K., and Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. *arXiv preprint arXiv:1009.6119*.
- Power, D. J. and Power, M. L. (2015). Sharing and analyzing data to reduce insurance fraud.
- Powers, D. M. (2011). Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation.

- Prekopcsak, Z., Makrai, G., Henk, T., and Gaspar-Papanek, C. (2011). Radoop: Analyzing big data with rapidminer and hadoop. In *Proceedings of the 2nd RapidMiner community meeting and conference (RCOMM 2011)*, pages 1–12.
- Pulizzi, J. and Heandley, A. (2014). B2C content marketing: 2014. *Insurance Journal*.
- Punitha, N. and Amsaveni, R. (2011). Methods and techniques to protect the privacy information in privacy preservation data mining. *IJCTA/ NOV-DEC*.
- PWC and Metcalfe, B. (2014). Strategic and emerging trends in insurance markets in South Africa, Kenya and Nigeria. *Africa insurance trends*.
- Questier, F., Put, R., Coomans, D., Walczak, B., and Vander Heyden, Y. (2005). The use of cart and multivariate regression trees for supervised and unsupervised feature selection. *Chemometrics and Intelligent Laboratory Systems*, 76(1):45–54.
- Quiggle, J. (2017). Mob associate uses deer parts to stage car wrecks. [Online] Available from: <http://www.insurancefraud.org/article.htm?RecID=3497>. [Accessed: 01/08/2017].
- Quinlan, J. R. (1993). C 4.5: Programs for machine learning. *The Morgan Kaufmann Series in Machine Learning, San Mateo, CA: Morgan Kaufmann, c1993*.
- Quinlan, J. R. (2014). *C4. 5: programs for machine learning*. Elsevier.
- R Core Team (2000). R language definition. *Vienna, Austria: R foundation for statistical computing*.
- Raghupathi, W. and Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*, 2(1):3.
- Rajesh, N., Sujatha, K., and Lawrence, A. A. (2016). Survey on privacy preserving data mining techniques using recent algorithms. *International Journal of Computer Applications*, 133(7):30–3.
- Rangra, K. and Bansal, K. (2014). Comparative study of data mining tools. *International journal of advanced research in computer science and software engineering*, 4(6).
- Rathna, S. S. and Karthikeyan, T. (2015). Survey on recent algorithms for privacy preserving data mining. *International Journal of Computer Science and Information Technologies*, 6(2):1835–40.
- Rawte, V. and Anuradha, G. (2015). Fraud detection in health insurance using data mining techniques. In *Communication, Information & Computing Technology (IC-CICT), 2015 International Conference on*, pages 1–5. IEEE.
- Red Edge Solutions (2015). The 8 compliance conditions. [Online] Available from: <http://www.popi360solution.co.za/the-8-compliance-details/>. [Accessed: 20/10/2017].

- Rish, I. (2001). An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46. IBM New York.
- Rong, Z., Xia, D., and Zhang, Z. (2013). Complex statistical analysis of big data: implementation and application of apriori and FP-growth algorithm based on mapreduce. In *Software Engineering and Service Science (ICSESS), 2013 4th IEEE International Conference on*, pages 968–972. IEEE.
- Rudolph, M. J. and MAAA, F. C. C. (2011). Us insurance company investment strategies in an economic downturn. *White Paper sponsored by Committee on Financial Research, Society of Actuaries (Schaumburg, IL)*.
- Russom, P. et al. (2011). Big data analytics. *TDWI Best Practices Report, Fourth Quarter*, pages 1–35.
- Sadiq, S., Orłowska, M., Sadiq, W., and Foulger, C. (2004). Data flow and validation in workflow modelling. In *Proceedings of the 15th Australasian database conference-Volume 27*, pages 207–214. Australian Computer Society, Inc.
- Sagioglu, S. and Sinanc, D. (2013). Big data: A review. In *Collaboration Technologies and Systems (CTS), 2013 International Conference on*, pages 42–47. IEEE.
- Sahin, Y., Bulkan, S., and Duman, E. (2013). A cost-sensitive decision tree approach for fraud detection. *Expert Systems with Applications*, 40(15):5916–5923.
- Sahl Andersen, J., De Fine Olivarius, N., and Krasnik, A. (2011). The danish national health service register. *Scandinavian Journal of Public Health*, 39(7_suppl):34–37.
- SAICB (2012). The South African insurance crime bureau. <http://www.saicb.co.za/>.
- SAICB (2013). June 2013 Case history. [Online] Available from: <http://www.saicb.co.za/resources/case-histories/129-vehicle-finance-fraud-update-wildfire-saicb-case-update>. [Accessed: 01/08/2017].
- SAICB (2014). February 2014 Case history. [Online] Available from: <http://www.saicb.co.za/resources/case-histories/150-project-loskop-insurance-fraud-cases>. [Accessed: 01/08/2017].
- Salloum, S., Huang, J. Z., and He, Y. (2016). Empirical analysis of asymptotic ensemble learning for big data. In *Proceedings of the 3rd IEEE/ACM International Conference on Big Data Computing, Applications and Technologies*, pages 8–17. ACM.
- Saloustros, G. and Magoutis, K. (2015). Rethinking hbase: Design and implementation of an elastic key-value store over log-structured local volumes. In *Parallel and Distributed Computing (ISPDC), 2015 14th International Symposium on*, pages 225–234. IEEE.
- Santam (2017). This is the santam group. [Online] Available from: <https://www.santam.co.za/financial/integrated-report/business-review/this-is-the-santam-group/>. [Accessed: 22/09/2017].

- Saporito, P. (2015). *Applied Insurance Analytics: A Framework for Driving More Value from Data Assets, Technologies, and Tools*. Financial Times Press.
- Sasaki, Y. et al. (2007). The truth of the f-measure. *Teach Tutor mater*, 1(5).
- Sathya, R. and Abraham, A. (2013). Comparison of supervised and unsupervised learning algorithms for pattern classification. *International Journal of Advanced Research in Artificial Intelligence*, 2(2):34–38.
- Sato, K. (2012). An inside look at google bigquery. *White paper*, URL: <https://cloud.google.com/files/BigQueryTechnicalWP.pdf>.
- Schiuma, G., Gavrilova, T., and Andreeva, T. (2012). Knowledge elicitation techniques in a knowledge management context. *Journal of Knowledge Management*, 16(4):523–537.
- Schneiberg, M. and Bartley, T. (2001). Regulating American industries: Markets, politics, and the institutional determinants of fire insurance regulation. *American Journal of Sociology*, 107(1):101–146.
- Sera, K. (2016). Tis the season for insurance fraud. [Online] Available from: <http://www.itweb.co.za/index.php?option=comcontentview=articleid=148940>. [Accessed: 30/03/2016].
- Shanahan, J. G. and Dai, L. (2015). Large scale distributed data science using apache spark. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2323–2324. ACM.
- Sharma, A. and Panigrahi, P. K. (2013). A review of financial accounting fraud detection based on data mining techniques. *arXiv preprint arXiv:1309.3944*.
- Shi, Y., Sun, C., Li, Q., Cui, L., Yu, H., and Miao, C. (2016). A fraud resilient medical insurance claim system. In *AAAI*, pages 4393–4394.
- Sibanda, W. and Pretorius, P. (2012). Comparative study of the application of box behnken design (bbd) and binary logistic regression (blr) to study the effect of demographic characteristics on hiv risk in South Africa. *Journal of Applied Medical Sciences*, 1(2):15–40.
- Siegel, E. (2016). *Predictive analytics: The power to predict who will click, buy, lie, or die*. Wiley Hoboken (NJ).
- Siegelman, P. (2014). Information & equilibrium in insurance markets with big data. *Conn. Ins. LJ*, 21:317.
- Simon-Tuval, T., Horev, T., and Kaplan, G. (2015). Medical loss ratio as a potential regulatory tool in the israeli healthcare system. *Israel journal of health policy research*, 4(1):21.

- Singhal, S. and Jena, M. (2013). A study on weka tool for data preprocessing, classification and clustering. *International Journal of Innovative technology and exploring engineering (IJITEE)*, 2(6):250–253.
- Singleton, T. W., Singleton, A. J., Bologna, G. J., and Lindquist, R. J. (2006). *Fraud auditing and forensic accounting*. John Wiley & Sons.
- Sithic, H. L. and Balasubramanian, T. (2013). Survey of insurance fraud detection using data mining techniques. *arXiv preprint arXiv:1309.0806*.
- Slack, E. (2012). What is big data? [Online] Available from: <http://www.storageswitzerland.com/Articles/Entries/2012/8/3WhatisBigData.html>. [Accessed: 02/10/2017].
- Snae, C. (2007). A comparison and analysis of name matching algorithms. *International Journal of Applied Science. Engineering and Technology*, 4(1):252–257.
- sparklyr (2016). sparklyr: R interface for apache spark. [Online] Available from: <http://spark.rstudio.com/>. [Accessed: 17/07/2017].
- Sperandei, S. (2014). Understanding logistic regression analysis. *Biochemia medica: Biochemia medica*, 24(1):12–18.
- Srivastava, U. and Gopalkrishnan, S. (2015). Impact of big data analytics on banking sector: Learning for Indian banks. *Procedia Computer Science*, 50:643–652.
- Stamford, C. (2010). 10 Technologies with biggest impact on p/c insurance companies. *Gartner*.
- Staples, S. (2011). Insurance analytics: Going beyond risk. *Insurance Networking*.
- Steen, H. (2016). R server 2016 (8.0.5) installation for linux systems. [Online] Available from: <https://docs.microsoft.com/en-us/machine-learning-server/install/r-server-install-linux-server-805>. [Accessed: 28/09/2016].
- Steinberg, D. and Colla, P. (2009). Cart: classification and regression trees. *The top ten algorithms in data mining*, 9:179.
- Stephens, O. (2017). "installation instructions". [Online] Available from: <https://github.com/OpenRefine/OpenRefine/wiki/Installation-Instructions>. [Accessed: 30/03/2017].
- Steynberg, M. (2016). 10 things you need to know to become a data scientist. [Online] Available from: <https://www.sqlshack.com/10-things-need-know-become-data-scientist/>. [Accessed: 18/07/2017].
- Stone, C. J., Friedman, J., Breiman, L., and Olshen, R. (1984). Classification and regression trees. *Wadsworth International Group*, 8:452–456.
- Sturm, A. and Shehory, O. (2004). A framework for evaluating agent-oriented methodologies. In *Agent-Oriented Information Systems*, pages 94–109. Springer.

- Surbhi, S. (2017). Difference between life insurance and general insurance. [Online] Available from: <http://keydifferences.com/difference-between-life-insurance-and-general-insurance.html>. [Accessed: 12/10/2017].
- Swedloff, R. (2014). Risk classification's big data (r) evolution. *Conn. Ins. LJ*, 21:339.
- Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570.
- Swire, P. (1997). Markets, self-regulation, and government enforcement in the protection of personal information, in privacy and self-regulation in the information age by the us department of commerce. *NTIA*.
- Talesh, S. A. (2015). Insurance and the law. *International Encyclopedia of the Social & Behavioral Sciences*, 11.
- Tappenden, A., Huynh, T., Miller, J., Geras, A., and Smith, M. (2009). Agile development of secure web-based applications. *International Journal of Information Technology and Web Engineering*, 1(2).
- Taric, G. J. and Poovammal, E. (2017). A survey on privacy preserving data mining techniques. *Indian Journal of Science and Technology*, 8(1).
- TDI (2016). Organized criminal activity, 1st degree felony. [Online] Available from: <https://www.tdi.texas.gov/fraud/cases.html>. [Accessed: 01/08/2017].
- Tennyson, S. (2002). Insurance experience and consumers' attitudes toward insurance fraud. *Journal of Insurance Regulation*, 21(2):35.
- Tennyson, S. (2008). Moral, social, and economic dimensions of insurance claims fraud. *Social Research*, 75(4):1181–1204.
- Thomas, S. S. (2017). Efficiency of clustering techniques in unsupervised learning scenario for health insurance fraud detection.
- Tigani, J. and Naidu, S. (2014). *Google BigQuery Analytics*. John Wiley & Sons.
- Timofeev, R. (2004). Classification and regression trees (cart) theory and applications. *Humboldt University, Berlin*.
- Todd, J. D., Welch, S. T., Welch, O. J., and Holmes, S. A. (1999). Insurer vs. insurance fraud: Characteristics and detection. *Journal of Insurance Issues*, pages 103–124.
- Tomlinson, P. (2015). The RDR paradox : practice management. *MoneyMarketing*.
- Trowbridge, E. and Rose, A. (2015). Top insurance industry issues in 2015. [Online] Available from: <http://www.pwc.com/us/en/insurance/publications/assets/pwc-top-issues-the-insurance-industry-2015.pdf>. [Accessed: 08/02/2017].
- Trowbridge, E. and Rose, A. (2017). Top issues: An annual report. [Online] Available from: <https://www.pwc.com/us/en/insurance/publications/assets/pwc-2017-insurance-top-issues.pdf>. [Accessed: 16/02/2018].

- Tu, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of clinical epidemiology*, 49(11):1225–1231.
- USAA (2017). Insurance claims fraud. [Online] Available from: <https://www.usaa.com/inet/wc/insuranceclaimsfraudmainakredirecttrue>. [Accessed: 10/03/2017].
- Valentini, G. L., Lassonde, W., Khan, S. U., Min-Allah, N., Madani, S. A., Li, J., Zhang, L., Wang, L., Ghani, N., Kolodziej, J., et al. (2013). An overview of energy efficiency techniques in cluster computing systems. *Cluster Computing*, pages 1–13.
- van Jaarsveld, J., Mostert, F., and Mostert, J. (2015). The claims handling process of liability insurance in south africa. *Risk Governance and Control: Financial markets and institutions*, page 133.
- Van Lamsweerde, A. (2003). From system goals to software architecture. In *International School on Formal Methods for the Design of Computer, Communication and Software Systems*, pages 25–43. Springer.
- Van Wolferen, J., Inbar, Y., and Zeelenberg, M. (2013). Moral hazard in the insurance industry. *Netspar Panel Paper*, 33:1–76.
- Vapnik, V. and Chervonenkis, A. Y. (1964). A class of algorithms for pattern recognition learning. *Avtomat. i Telemekh*, 25(6):937–945.
- Vassiliadis, P., Simitsis, A., and Skiadopoulou, S. (2002). Conceptual modeling for etl processes. In *Proceedings of the 5th ACM international workshop on Data Warehousing and OLAP*, pages 14–21. ACM.
- Viaene, S. and Dedene, G. (2004). Insurance fraud: issues and challenges. *The Geneva Papers on Risk and Insurance-Issues and Practice*, 29(2):313–333.
- Von Solms, R. and Van Niekerk, J. (2013). From information security to cyber security. *Computers and security*, 38:97–102.
- Walker, M. A. (2015). The professionalisation of data science. *International Journal of Data Science*, 1(1):7–16.
- Waller, M. A. and Fawcett, S. E. (2013). Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management. *Journal of Business Logistics*, 34(2):77–84.
- Wang, J. and Gu, L. (2016). Challenges of teaching data science in a business school. *Issues in Information Systems*, 17(3).
- Wang, P. (2007). The logic of intelligence. In *Artificial General Intelligence*, pages 31–62. Springer.

- Wang, S.-L., Pai, H.-T., Wu, M.-F., Wu, F., and Li, C.-L. (2017). The evaluation of trustworthiness to identify health insurance fraud in dentistry. *Artificial Intelligence in Medicine*, 75:40–50.
- Wang, Y., Guo, C., and Song, L. (2009). Architecture of e-commerce systems based on j2ee and mvc pattern. In *Management of e-Commerce and e-Government, 2009. ICMECG'09. International Conference on*, pages 284–287. IEEE.
- Wei, W., Li, J., Cao, L., Ou, Y., and Chen, J. (2013). Effective detection of sophisticated online banking fraud on extremely imbalanced data. *World Wide Web*, pages 1–27.
- Wells, J. (2013). *Insurance fraud casebook : paying a premium for crime*. John Wiley & Sons Publishing.
- Werner, G. and Modlin, C. (2010). *Basic Ratemaking*. Casualty Actuarial Society, fourth edition edition.
- West, J. and Bhattacharya, M. (2016). Intelligent financial fraud detection: a comprehensive review. *Computers & Security*, 57:47–66.
- West, J., Bhattacharya, M., and Islam, R. (2014). Intelligent financial fraud detection practices: An investigation. In *International Conference on Security and Privacy in Communication Systems*, pages 186–203. Springer.
- Westreich, D., Lessler, J., and Funk, M. J. (2010). Propensity score estimation: neural networks, support vector machines, decision trees (cart), and meta-classifiers as alternatives to logistic regression. *Journal of clinical epidemiology*, 63(8):826–833.
- Wildemuth, B. M. (2004). The effects of domain knowledge on search tactic formulation. *Journal of the Association for Information Science and Technology*, 55(3):246–258.
- Wimmer, H. and Powell, L. M. (2016). A comparison of open source tools for data science. *Journal of Information Systems Applied Research*, 9(2):4.
- Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.
- Woods, D. (2014). Why google capital placed its hadoop bet on mapr. [Online] Available from: <https://www.forbes.com/sites/danwoods/2014/06/30/why-google-capital-placed-its-hadoop-bet-on-mapr/4d89096778e7>. [Accessed: 19/07/2017].
- Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Philip, S. Y., et al. (2008). Top 10 algorithms in data mining. *Knowledge and information systems*, 14(1):1–37.
- Xu, L., Jiang, C., Chen, Y., Wang, J., and Ren, Y. (2016). A framework for categorizing and applying privacy-preservation techniques in big data mining. *Computer*, 49(2):54–62.

- Xu, L., Jiang, C., Wang, J., Yuan, J., and Ren, Y. (2014). Information security in big data: privacy and data mining. *IEEE Access*, 2:1149–1176.
- Xu, W., Wang, S., Zhang, D., and Yang, B. (2011). Random rough subspace based neural network ensemble for insurance fraud detection. In *Computational Sciences and Optimization (CSO), 2011 Fourth International Joint Conference on*, pages 1276–1280. IEEE.
- Yang, T. (2013). *Package ‘RHive’*, 0.07 edition.
- Yoon, K., Hoogduin, L., and Zhang, L. (2015). Big data as complementary audit evidence. *Accounting Horizons*, 29(2):431–438.
- Young, T. (2007). Manageability, Maintainability, and Supportability. [Online] Available from: <https://msdn.microsoft.com/en-us/library/bb896744.aspx>. [Accessed: 15/08/2017].
- Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauley, M., Franklin, M. J., Shenker, S., and Stoica, I. (2012). Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*, pages 2–2. USENIX Association.
- Zareapoor, M. and Shamsolmoali, P. (2015). Application of credit card fraud detection: Based on bagging ensemble classifier. *Procedia Computer Science*, 48:679–685.
- Zhao, Y. and Bhowmick, S. S. (2015). Association rule mining with r. *A Survey Nanyang Technological University, Singapore*.
- Zheng, J. and Dagnino, A. (2014). An initial study of predictive machine learning analytics on large volumes of historical data for power system applications. In *Big Data (Big Data), 2014 IEEE International Conference on*, pages 952–959. IEEE.
- Zhu, J., Zou, H., Rosset, S., and Hastie, T. (2009). Multi-class adaboost. *Statistics and its Interface*, 2(3):349–360.
- Zikopoulos, P., Eaton, C., et al. (2011). *Understanding big data: Analytics for enterprise class hadoop and streaming data*. McGraw-Hill Osborne Media.

Appendix A

Example Code used to Predict Fraud

```
1 public class FraudPredictor
2 {
3     public Claim claim;
4     public DataMasker dataMasker;
5     public DataFormatter dataFormatter;
6     public FraudulentClaimLogistRegression
7         fraudulentClaimLogistRegression;
8     public FraudulentClaimSupervisedRules
9         fraudulentClaimSupervisedRules;
10    public FraudulentClaimUnSupervisedRules
11        fraudulentClaimUnSupervisedRules;
12    public ClaimTotalFraudProbability CTFP;
13
14
15
16
17    public FraudPredictor()
18    {
19
20    }
21
22    public void addClaim(Claim newClaim)
23    {
24        claim = newClaim;
25    }
26
27    public string getFraudReasoning()
28    {
29        return "The total fraud probabilty is:" + CTFP * 100 + "%";
30    }
31
32    public bool isFraudulent()
33    {
34        //Claim is masked
35        claim = dataMasker.MaskData(claim);
36        //Claim is formatted
37        claim = dataFormatter.FormatData(claim);
38    }
39 }
```

```
32 //Add Logistic Regression Probability To Total Fraud
    Probability
33 CTFP.AddLogisticRegression(fraudulentClaimLogistRegression.
    GetLogisticRegression(claim));
34 //For each supervised rule that is triggered, add the SAW
    to the Total Fraud Probability
35 foreach(Rule BrokenRule in fraudulentClaimSupervisedRules.
    GetFiredRules(claim))
36 {
37     CTFP.IncreaseNumberOFSupervisedRulesFired();
38 }
39 //For each un supervised rule that is broken, add the RBVP
    to the Total Fraud Probability
40 foreach (Rule BrokenRule in
    fraudulentClaimUnSupervisedRules.GetFiredRules(claim))
41 {
42     CTFP.IncreaseNumberOFUnSupervisedRulesBroken();
43 }
44
45 //If the CTFP is greater than the FPT, declare that the
    claim is fraudulent
46 if (CTFP.TotalProbability() >= Maintenance.FPT)
47 {
48     return true;
49 }
50 else
51 {
52     return false;
53 }
54
55 }
56
57
58 }
59
60 public class testClaims
61 {
62     static void Main()
63     {
64         // New Claim is added by service consultant
65         Claim claimSubmittedBySC = new Claim();
66         //New Fraud Predictor is Generated and claim is added to it
67         FraudPredictor fraudpredictor = new FraudPredictor();
68         fraudpredictor.addClaim();
69
70         //Check to see if claim is fraudulent
71         if(fraudpredictor.isFraudulent())
72         {
73             claimSubmittedBySC.PayClaim();
```

```
74     }
75     else
76     {
77         claimSubmittedBySC.RejectClaim();
78         MessageBox.show(fraudpredictor.getFraudReasoning());
79     }
80
81     //Add claim to training data-set
82     //This must only be called once the claim has been
83     //investigated if isFraudulent is true
84     claimSubmittedBySC.AddToTrainingDataSet();
85
86 }
87 }
```

Appendix B

SQL Code used to Generate Logistic Regression Test Data

```
1
2 CREATE TABLE #Ages
3 (Age int)
4
5 DECLARE @startnum INT, @endnum INT
6
7 SET @startnum =10
8 SET @endnum =110
9
10 ;
11 WITH gen AS (
12     SELECT @startnum AS num
13     UNION ALL
14     SELECT num+10 FROM gen WHERE num+10<=@endnum
15 )
16 INSERT INTO #Ages (Age)
17 SELECT * FROM gen
18 option (maxrecursion 10000)
19
20
21
22 CREATE TABLE #SumInsureds
23 (SumInsured int)
24
25 SET @startnum =10000
```

```

26 SET @endnum =5000000
27
28 ;
29 WITH gen AS (
30     SELECT @startnum AS num
31     UNION ALL
32     SELECT num+50000 FROM gen WHERE num+50000<=@endnum
33 )
34 INSERT INTO #SumInsureds (SumInsured)
35 SELECT * FROM gen
36 option (maxrecursion 10000)
37
38 CREATE TABLE #Date_diffs
39 (Date_diff int)
40
41 SET @startnum =0
42 SET @endnum =6000
43
44 ;
45 WITH gen AS (
46     SELECT @startnum AS num
47     UNION ALL
48     SELECT num+10 FROM gen WHERE num+10<=@endnum
49 )
50 INSERT INTO #Date_diffs (Date_diff)
51 SELECT * FROM gen
52 option (maxrecursion 10000)
53
54 CREATE TABLE #AmountPaiids
55 (AmountPaid int)
56
57
58 SET @startnum =10000
59 SET @endnum =5000000
60
61 ;
62 WITH gen AS (
63     SELECT @startnum AS num
64     UNION ALL
65     SELECT num+50000 FROM gen WHERE num+50000<=@endnum
66 )
67 INSERT INTO #AmountPaiids (AmountPaid)
68 SELECT * FROM gen
69 option (maxrecursion 10000)
70
71 CREATE TABLE #TotalPoliciesRevenues
72 (TotalPoliciesRevenue int)
73
74

```



```
75 SET @startnum =0
76 SET @endnum =5000000
77
78 ;
79 WITH gen AS (
80     SELECT @startnum AS num
81     UNION ALL
82     SELECT num+10000 FROM gen WHERE num+10000<=@endnum
83 )
84 INSERT INTO #TotalPoliciesRevenues (TotalPoliciesRevenue)
85 SELECT * FROM gen
86 option (maxrecursion 10000)
87
88 SELECT *, (ABS(CHECKSUM(NewId())) % 6)*SumInsured*0.1 AS Excess
89 FROM #AmountPays
90 CROSS JOIN #Ages
91 CROSS JOIN #TotalPoliciesRevenues
92 CROSS JOIN #Date_diffs
93 CROSS JOIN #SumInsureds
94
95 DROP TABLE #TotalPoliciesRevenues, #Date_diffs, #SumInsureds, #
    AmountPays,#Ages
```

Appendix C

Unsupervised Machine Learning Rules

Refer to <https://tinyurl.com/UPInsurClaimsFraud> sub-folder: 6. Testing\1. Rules\Unsupervised Rules.xlsx to view unsupervised machine learning rules.

Appendix D

Supervised Machine Learning Rules

Refer to <https://tinyurl.com/UPInsurClaimsFraud> sub-folder: 6. Testing\1. Rules\Supervised Rules.xlsx to view supervised machine learning rules.

Appendix E

OpenRefine API Call to Remove Empty Rows

```
1 [
2   {
3     "op": "core/row-removal",
4     "description": "Remove rows",
5     "engineConfig": {
6       "mode": "row-based",
7       "facets": [
8         {
9           "omitError": false,
10          "expression": "isBlank(value)",
11          "selectBlank": false,
12          "selection": [
13            {
14              "v": {
15                "v": true,
16                "l": "true"
17              }
18            }
19          ],
20          "selectError": false,
21          "invert": false,
22          "name": "ID",
23          "omitBlank": false,
24          "type": "list",
25          "columnName": "ID"
26        }
27      ]
28    }
29  }
30 ]
```

Appendix F

OpenRefine API Call to Confirm DateOfClaim is a Date Value

```
1 [
2   {
3     "op": "core/text-transform",
4     "description": "Text transform on cells in column DateOfClaim
5       using expression grel:value.toDate()",
6     "engineConfig": {
7       "mode": "row-based",
8       "facets": []
9     },
10    "columnName": "DateOfClaim",
```

```
10     "expression": "grel:value.toDate()",
11     "onError": "keep-original",
12     "repeat": false,
13     "repeatCount": 10
14 },
15 {
16     "op": "core/row-removal",
17     "description": "Remove rows",
18     "engineConfig": {
19         "mode": "row-based",
20         "facets": [
21             {
22                 "selectNonTime": true,
23                 "expression": "value",
24                 "selectBlank": true,
25                 "selectError": true,
26                 "selectTime": false,
27                 "name": "DateOfClaim",
28                 "from": 1167948000000,
29                 "to": 1488837600000,
30                 "type": "timerange",
31                 "columnName": "DateOfClaim"
32             }
33         ]
34     }
35 }
36 ]
```

Appendix G

OpenRefine API Call to Confirm AmountPaid is a Numeric Value

```
1 [
2   {
3     "op": "core/text-transform",
4     "description": "Text transform on cells in column AmountPaid using
5       expression grel:value.toNumber()",
6     "engineConfig": {
7       "mode": "row-based",
8       "facets": []
9     },
10    "columnName": "AmountPaid",
11    "expression": "grel:value.toNumber()",
12    "onError": "keep-original",
13    "repeat": false,
```

```
13     "repeatCount": 10
14   },
15   {
16     "op": "core/row-removal",
17     "description": "Remove rows",
18     "engineConfig": {
19       "mode": "row-based",
20       "facets": [
21         {
22           "selectNumeric": false,
23           "expression": "value",
24           "selectBlank": true,
25           "selectNonNumeric": true,
26           "selectError": true,
27           "name": "AmountPaid",
28           "from": 0,
29           "to": 130000000,
30           "type": "range",
31           "columnName": "AmountPaid"
32         }
33       ]
34     }
35   }
36 ]
```

Appendix H

OpenRefine API Call to Confirm PolicyStartDate is Before DateOfClaim

```
1 [
2   {
3     "op": "core/row-removal",
4     "description": "Remove rows",
5     "engineConfig": {
6       "mode": "row-based",
7       "facets": [
8         {
9           "selectNumeric": true,
10          "expression": "grel:diff(now(), cells['PolicyStartDate'].
11              value, \"days\")",
12          "selectBlank": true,
13          "selectNonNumeric": true,
14          "selectError": true,
15          "name": "DateOfClaim",
16          "from": -1000000,
```

```
16         "to": 0,  
17         "type": "range",  
18         "columnName": "DateOfClaim"  
19     }  
20 ]  
21 }  
22 }  
23 ]
```

Appendix I

OpenRefine API Call to Confirm ExcessPaid is Greater than Zero

```
1  [  
2  {  
3      "op": "core/text-transform",  
4      "description": "Text transform on cells in column ExcessPaid using  
5          expression grel:value.toNumber()",  
6      "engineConfig": {  
7          "mode": "row-based",  
8          "facets": []  
9      },  
10     "columnName": "ExcessPaid",  
11     "expression": "grel:value.toNumber()",  
12     "onError": "keep-original",  
13     "repeat": false ,  
14     "repeatCount": 10  
15 },  
16 {  
17     "op": "core/row-removal",  
18     "description": "Remove rows",  
19     "engineConfig": {  
20         "mode": "row-based",  
21         "facets": [  
22             {  
23                 "selectNumeric": true ,  
24                 "expression": "value",  
25                 "selectBlank": true ,  
26                 "selectNonNumeric": false ,  
27                 "selectError": true ,  
28                 "name": "ExcessPaid",  
29                 "from": -100000,  
30                 "to": 0,  
31                 "type": "range",  
32                 "columnName": "ExcessPaid"
```

32 }

33] }

34 } }

35 } }

36] }

Appendix J

Architecture Capacity Testing Results

Table 1: Table Showing Results of Capacity Testing

Number of Claim Records	Time Elapsed to Train IFAMLM (s)	Time Elapsed to Train IFLMLM (s)
20,000	4.46	1.94
40,000	7.20	3.30
60,000	9.14	4.27
80,000	11.39	5.76
100,000	13.62	6.92
120,000	15.32	7.76
140,000	17.78	8.89
160,000	19.86	9.80
180,000	21.45	10.70
200,000	22.86	11.70
220,000	24.83	13.03
240,000	26.76	13.38
260,000	29.63	14.60
280,000	31.69	14.94
300,000	33.22	16.26
320,000	35.45	16.66
340,000	37.09	18.03
360,000	38.50	19.10
380,000	40.72	19.76
400,000	42.66	20.72
420,000	44.31	21.51
440,000	46.74	23.76
460,000	47.49	23.94
480,000	50.62	24.53
500,000	52.64	25.35