View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Sydney eScholarship

DOCTORAL THESIS

FIGHTING ACCOUNTING FRAUD THROUGH FORENSIC ANALYTICS

Author:
Maria JOFRE

Supervisor:
Prof. Richard GERLACH
Dr. Marcel SCHARTH

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

in the

Discipline of Business Analytics
Business School
The University of Sydney

July 2017

Declaration of Authorship

I, Maria JOFRE, declare that this thesis, titled "Fighting Accounting Fraud Through Forensic Analytics" and the work presented in it are my own. This thesis has not been submitted for any degree or other qualification at this University or any other institution.

I confirm that the intellectual content of this thesis is the product of my own work, that I have quoted all published work of others and that I have acknowledged all main sources of help.

Maria Jofre

Very few of the common people realise that
the political and legal systems
have been corrupted by decades
of corporate lobbying.

Steven Magee

Abstract

Accounting Fraud is one of the most harmful financial crimes as it often results in massive corporate collapses, commonly silenced by powerful high-status executives and managers. Accounting fraud represents a significant threat to the financial system stability due to the resulting diminishing of the market confidence and trust of regulatory authorities. Its catastrophic consequences expose how vulnerable and unprotected the community is in regards to this matter, since most damage is inflicted to investors, employees, customers and government.

Accounting fraud is defined as the calculated misrepresentation of the financial statement information disclosed by a company in order to mislead stakeholders regarding the firm?s true financial position. Different fraudulent tricks can be used to commit accounting fraud, either direct manipulation of financial items or creative methods of accounting, hence the need for non-static regulatory interventions that take into account different fraudulent patterns. Accordingly, this study aims to identify signs of accounting fraud occurrence to be used to, first, identify companies that are more likely to be manipulating financial statement reports, and second, assist the task of examination within the riskier firms by evaluating relevant financial red-flags, as to efficiently recognise irregular accounting malpractices. To achieve this, a thorough forensic data analytic approach is proposed that includes all pertinent steps of a data-driven methodology.

First, data collection and preparation is required to present pertinent information related to fraud offences and financial statements. The compiled sample of known fraudulent companies is identified considering all Accounting Series Releases and Accounting and Auditing Enforcement Releases issued by the U.S. Securities and Exchange Commission between 1990 and 2012, procedure that resulted in 1,594 fraud-year observations. Then, an in-depth financial ratio analysis is performed in order to evaluate publicly available financial statement data and to preserve only meaningful predictors of accounting fraud. In particular, two commonly used statistical approaches, including non-parametric hypothesis testing and correlation analysis, are proposed to assess significant differences between corrupted and genuine reports as well as to identify associations between the considered ratios. The selection of a smaller subset of explanatory variables is later reinforced by the implementation of a complete subset logistic regression methodology.

Finally, statistical modelling of fraudulent and non-fraudulent instances is performed by implementing several machine learning methods. Classical classifiers are considered first as benchmark frameworks, including logistic regression and discriminant analysis. More complex techniques are implemented next based on decision trees bagging and boosting, including bagged trees, AdaBoost and random forests.

In general, it can be said that a clear enhancement in the understanding of the fraud phenomenon is achieved by the implementation of financial ratio analysis, mainly due to the interesting exposure of distinctive characteristics of falsified reporting and the selection of meaningful ratios as predictors of accounting fraud, later validated using a combination of logistic regression models. Interestingly, using only significant explanatory variables leads to similar results obtained when no selection is performed. Furthermore, better performance is accomplished in some cases, which strongly evidences the convenience of employing less but significant information when detecting accounting fraud offences.

Moreover, out-of-sample results suggest there is a great potential in detecting falsified accounting records through statistical modelling and analysis of publicly available accounting information. It has been shown good performance of classic models used as benchmark and better performance of more advanced methods, which supports the usefulness of machine learning models as they appropriately meet the criteria of accuracy, interpretability and cost-efficiency required for a successful detection methodology.

This study contributes in the improvement of accounting fraud detection in several ways, including the collection of a comprehensive sample of fraud and non-fraud firms concerning all financial industries, an extensive analysis of financial information and significant differences between genuine and fraudulent reporting, selection of relevant predictors of accounting fraud, contingent analytical modelling for better differentiate between non-fraud and fraud cases, and identification of industry-specific indicators of falsified records.

The proposed methodology can be easily used by public auditors and regulatory agencies in order to assess the likelihood of accounting fraud and to be adopted in combination with the experience and instinct of experts to lead to better examination of accounting reports. In addition, the proposed methodological framework could be of assistance to many other interested parties, such as investors, creditors, financial and economic analysts, the stock exchange, law firms and to the banking system, amongst others.

Acknowledgements

First of all, I would like to thank my supervisory team who has patiently guided me during this fascinating journey. A big thank you to my main supervisor Prof. Richard Gerlach and to my co-supervisor Dr. Marcel Scharth for their constant support. They both have contributed significantly to my research and professional achievements. I would like to extend my acknowledge to Dr. Demetris Christodoulou who supported me during the first years of the program and helped me with all administrative processes and other issues. I am also indebted to the thesis examiners who have made meaningful corrections and suggestions, which have improved the quality of the work presented in this document.

Special thanks to Conicyt, institution responsible for providing financial support to Chilean students, as without the provided scholarship (Becas Chile) it would have been very difficult to afford my studies and living expenses during this 4-year period.

In addition, I would like to thank the Securities Class Action Clearinghouse, Stanford Law School, for providing access to a very comprehensive database of accounting fraud cases.

Most importantly, a huge thanks to my mom, dad and brother, who have given me unconditional love and support throughout my life. Especially grateful to my grandmother Maria Elizabeth, who lights candles for every little PhD-related event. Also, absolutely thankful to my late grandmother Alicia Berta de Jesus (R.I.P.), to whom I would love to hug and share a laugh, or maybe a tear or two.

Always grateful to all my friends who have had an influence on my life in some unique and magical way. I cannot express enough how eternally thankful I am to share my life with such exceptional beings.

Contents

List of Figures

List of Tables

List of Abbreviations

AAER	**A**ccounting and **A**uditing **E**nforcement **R**eleases
AB	**A**da**B**oost
ACFE	**A**ssociation of **C**ertified **F**rauf **E**xaminers
AF	**A**ccounting **F**raud
ASR	**A**ccounting **S**eries **R**eleases
AUC	**A**rea **U**nder the **C**urve
BT	**B**oosted **T**rees
CACL	**C**urrent **A**ssets to **C**urrent **L**iabilities
CART	**C**lassification **A**nd **R**egression **T**rees
CATA	**C**urrent **A**ssets to **T**otal **A**ssets
CFFONI	**C**ash **F**low **F**rom **O**perations to **N**et **I**ncome
CHNI	**C**as**H** to **N**et **I**ncome
CIK	**C**entral **I**ndex **K**ey number
CSLR	**C**omplete **S**ubset **L**ogistic **R**egression
CUSIP	**C**ommitee on **U**niform **S**ecurities **I**dentification **P**rocedures
DA	**D**iscriminant **A**nalysis
DT	**D**ecision **T**rees
EBITTA	**E**arnings **B**efore **I**nterest and **T**ax to **T**otal **A**ssets
GAAP	**G**enerally **A**ccepted **A**ccounting **P**rinciples
GVKEY	**G**lobal company **KEY**
IVCA	**I**n**V**entory to **C**urrent **A**ssets
IVCOGS	**I**n**V**entory to **C**ost **O**f **G**ood **S**old
IVSA	**I**n**V**entory to total **SA**les
IVTA	**I**n**V**entory to **T**otal **A**ssets
LDA	**L**inear **D**iscriminant **A**nalysis
LR	**L**ogistic **R**egression
LTDTA	**L**ong-**T**erm **D**ebt to **T**otal **A**ssets
NITA	**N**et **I**ncome to **T**otal **A**ssets

NNs	**N**eura **N**etwoks
PYCOGS	accounts **P**a**Y**able to **C**ost **O**f **G**ood **S**old
QDA	**Q**uadratic **D**iscriminant **A**nalysis
RETA	**R**etained **E**arnings to **T**otal **A**ssets
RF	**R**andom **F**orests
ROC	**R**eceiver **O**perating **C**haracteristic curve
RVSA	accounts **R**ecei**V**able to total **SA**les
RVTA	accounts **R**ecei**V**able to **T**otal **A**ssets
SATA	total **SA**les to **T**otal **A**ssets
SATE	total **SA**les to **T**otal **E**quity
SEC	U.S. **S**ecurity **E**xchange **C**ommission
SIC	**S**tandard **I**ndustrial **C**lassification system
TLTA	**T**otal **L**iabilities to **T**otal **A**ssets
TLTE	**T**otal **L**iabilities to **T**otal **E**quity
WCTA	**W**orking **C**apital to **T**otal **A**ssets

Dedicated to my parents,
Maria Angelica Alegria and
Fernando Jofre.
I love you deeply.

Chapter 1

Introduction

1.1 Motivation

In the last few decades, accounting fraud has been drawing a great deal of attention amongst researchers and practitioners, as it is becoming increasingly frequent and complex, and therefore, more difficult to prevent and control it effectively. It is considered to be one of the most harmful corporate crimes (Mokhiber and Weissman, 2005), since it is believed to be connected to other major white-collar crimes, such as securities offences, organised crime and money laundering.

In the framework of this thesis, accounting fraud is defined as the calculated misrepresentation of the financial statement information that is publicly disclosed by companies. The intention is to mislead stakeholders regarding the firm's true financial position, by overstating its expectations on assets, or understating exposure to liabilities; hence the artificial inflation of earnings, as well as its return on equity. Accounting fraud may take the form of either direct manipulation of financial items or via creative methods of accounting (Shilit and Perler, 2010).

The catastrophic consequences of accounting fraud evidence how vulnerable and unprotected the community is in regard to this matter, since most damage is inflicted to investors, employees, customers and regulatory authorities. Accounting fraud often results in massive corporate collapses and deterioration of market confidence (Ngai et al., 2011), which is rapidly silenced by powerful high-status executives and managers. Several accounting scandals reflect this reality, the Enron infamous case being one of the most controversial. Exposed in October 2001, this scam concluded with the bankruptcy of the company, followed by 4,500 employees who lost their jobs and pension funds, and an estimated loss of 74 billion dollars assumed by investors and stakeholders.

Perpetrators of accounting fraud can be motivated by personal benefit (e.g.: maximisation of compensation packages), or by explicit or implied contractual obligations such as debt covenants, and the need to meet market projections and expected economic growth. The most harm is inflicted to the long-run reputation of the organisation itself, the value destruction of investors and the diminishing of the public's trust in the capital market. Other victims often include employees, suppliers, partners, customers, regulatory institutions, enforcement agencies, taxation authorities, the stock exchange, creditors and financial analysts (Pai, Hsu, and Wang, 2011).

Standard auditing procedures are often insufficient to identify fraudulent accounting reports since most managers recognise the limitations of audits, hence the need for additional comprehensive analytical methods to detect accounting fraud accurately and in an early stage (Kaminski, Wetzel, and Guan, 2004). In addition, given their hidden dynamic characteristics, 'book cooking' accounting practices are particularly hard to detect, thus the importance of more sophisticated tools to be used to assist the early identification of risk signs and to further expose complex fraudulent schemes.

Although several data-informed quantitative models have been developed to automate and reduce the manual auditing processes related to false reporting identification (Bose, Piramuthu, and Shaw, 2011), these are not sufficient to uncover complex fraudulent structures and to identify warning signs of accounting fraud. Accordingly, the present study aims to improve the detection of accounting fraud offences through a detailed analysis of discovered fraudulent cases and an exhaustive evaluation of financial features that best determine the occurrence of falsified accounting reporting.

1.2 Problem Statement

The research question seeking to be answered by this thesis is defined as follows:

> *What financial red-flags should be examined in order to detect accounting fraud?*

To properly address the previous question, it is required to investigate the following key components:

1. Relevant financial information for accounting fraud detection.
2. Machine learning methods to estimate the likelihood of fraud occurrence.
3. Risk indicators for effective accounting fraud detection.

Accordingly, the main objective of this research project is to improve the detection rate of accounting fraud offences through the implementation of several machine learning methods and assessment of key financial risk indicators, in order to potentially assist the design of an innovative, flexible and responsive regulatory tool.

1.3 Research Approach

In order to achieve the proposed objective and to further answer the research question, a detailed methodology has been implemented considering all relevant steps of a forensic data analysis approach: (i) data collection, preparation and validation; (ii) examination of potential explanatory variables and further selection of the most relevant ones; (iii) modelling of the fraudulent phenomenon; (iv) critical assessment of statistical models; (v) interpretation of results; and (vi) conclusions and suggestions related to corporate regulation and examination of financial reports.

1.4 Contributions

The main contributions of this thesis are the following:

1. **Examination of the state of the art associated with accounting fraud detection:** Critical review of previous studies related to statistical methods and their capability of detecting fraudulent financial reports. Analysis of the proposed methodologies, sample selection process, chosen sample size, explanatory

variables considered, data mining models employed and achieved predictive accuracy.

2. **Collection of a relative large database of accounting fraud offences:** Manual compilation of accounting fraud cases of public U.S. companies and their corresponding financial statement data. This includes the information of all litigation releases related to accounting fraud offences published by the U.S. Security and Exchange Commission (SEC) between 1990 and 2012. This task is considered by the author as one of the most relevant, since a great effort was made to gather substantial information to be used as an adequate representation of the population of accounting fraud offences.

3. **Evaluation of financial information related to fraudulent reporting:** In depth analysis of financial reporting items in order to deduce the relationship between unethical behaviour and business performance. Moreover, an exhaustive assessment of relevant financial ratios has been performed to better understand illicit accounting practices across all industries, as well as within specific economic domains.

4. **Modelling of the accounting fraud phenomenon:** Implementation and assessment of several analytical models to first assist the selection of significant explanatory variables, and second to achieve satisfactory detection rates. What is more, the outcome of the proposed modelling exercise is considered of value, as interesting fraudulent behaviours are exposed that help with the identification of premature warning signs of accounting fraud.

5. **Recommendation of financial indicators for adaptive corporate regulation:** Identification of industry-specific financial ’red flags’ to be used for effective examination of public documents. Relevant financial indicators are suggested to be part of the regulatory agenda, in order to better support the control and prevention of accounting fraud offences, and to further strengthen the regulatory system in performing this task.

1.5 Thesis Outline

An overview of the chapters constituting this thesis is presented below:

Chapter 2: The Accounting Fraud Phenomenon

The main objective of this chapter is to attain a better understanding of the phenomenon of accounting fraud. In light of this, a preliminary theoretical framework is elaborated with regard to white-collar crime, corporate crime and accounting fraud. A brief explanation of some of the worst financial reporting scandals is given to evidence the tragic consequences of accounting fraud, followed by a description of potential victims. Then, after a careful review of the literature, it is argued the use of forensic accounting techniques to appropriately approach the phenomenon of accounting fraud. Finally, the proposed methodology is introduced and justified in the basis of what is missing or could be improved, and what can be done to enhance the detection of accounting fraud from an empirical point of view.

Chapter 3: Forensic Data Analysis

In this chapter, the focus is primarily the data. A complete detail of how the data was collected is presented in this section along with the description of the following steps of data cleaning, data transformation, data merging and data validation.

Chapter 4: Financial Ratio Analysis

The final goal of Chapter 4 is variable selection, that is, finding the most significant explanatory variables for detecting accounting fraud offences. It essentially supports the use of financial ratios as predictors of falsified reporting by means of analysing statistically significant differences between ratios obtained from fraudulent reports and genuine reports. This ratio analysis is first performed irrespectively of the economic sector, and then taking into consideration the industry division where companies belong to.

Chapter 5: Complete Subset Logistic Regression

An innovative statistical approach is presented and further expanded to support the number of explanatory variables selected in the previous chapter.

Chapter 6: Accounting Fraud Modelling

In this chapter, the objective is to obtain satisfactory detection rates of accounting fraud offences. In order to achieve this, several machine learning methods are described and further implemented for determining the likelihood of accounting fraud occurrence. Finally, an interesting interpretation of numerical results is performed again for all industries, and then by industry division.

Chapter 7: Financial Indicators of Accounting Fraud

Results from the previous chapter are expanded to identify industry-specific indicators of accounting fraud. Chapter 7 promotes the use of these financial red flags when oversight tasks are performed.

Chapter 8: Conclusions, Limitations and Future Work

In this final chapter, all steps of the proposed methodology are summarised as well as the main contributions of the study. The importance of understanding the phenomenon of accounting fraud, statistical modelling and further identification of relevant red-flags, is once again briefly explained and emphasised. Limitations of the suggested methodology are identified and explained, and future work is recommended based on the exhaustive data analysis performed throughout the thesis.

Chapter 2

The Accounting Fraud Phenomenon

In order to efficiently detect accounting fraud offences, a comprehensive understanding of the phenomenon is required. For this reason, a theoretical framework has to be elaborated first with regard to white-collar crime and corporate crime. Once the general context has been established, a complete overview of the phenomenon of accounting fraud is discussed to further evidence its catastrophic consequences.

2.1 Overview of White-Collar Crime

Almost 70 years ago, Edwin H. Sutherland introduced the term *white-collar crime* as 'a crime committed by a person of respectability and high social status in the course of his occupation' (Sutherland, 1949). Although Sutherland's definition has generated a great deal of criticism and controversy (Simpson and Weisburd, 2009), it exhibits the important issue of inappropriately recognise, prevent and control crimes perpetrated by persons in position of power (Benson and Simpson, 2014). The problem of white-collar crime is particularly serious and complex, since it occurs in almost every country and industry.

How white-collar offences are committed is specially intriguing, since it typically involves a professional using his or her expert knowledge to take advantage of people who trust that the professional will act on behalf of their best interests. In light of this, white-collar crime is an offence fully determined by the abuse of power of professionals who most likely perceive their deceptive behaviour as normal business practices, which is usually referred to as *moral insensitivity*.

As Benson and Simpson (2014) explain, white-collar crimes are always associated with criminal opportunities, which basically consist of a suitable target and lack of guardianship. Consequently, holding particular occupational positions will greatly facilitate access to white-collar crime opportunities. There are many types of white-collar crime, including fraud, antitrust violations, bribery, embezzlement, money laundering, environmental crimes and workplace crimes, amongst others. As such, different criminal opportunities and behaviours can be identified and further analysed in order to properly define deterrence strategies.

Many of these offences are committed by corporations or, in other words, by individuals acting on behalf of a corporation. It is believed that organisational crimes are significantly more expensive compared to individual offences basically due to the economic and political power of people in charge (Benson and Simpson, 2014). Fraudulent corporations often use their economic advantages and resources to hide illicit activities behind legitimate business, and further manipulate the legal environment within which they operate. A more detailed analysis of corporate crime is given in the following section.

2.2 Overview of Corporate Crime

As mentioned earlier, corporate crimes are basically white-collar crime offences committed by corporations or people acting on behalf of them. It includes all corporate activities that are proscribed and punishable by law (Braithwaite, 1984). In this case, what is pursued are organisational benefits rather than individual gains (Simpson, 2002). Many corporate crime offences are committed within large and complex organisations. As such, many people may be involved, making it very difficult to determine who is responsible for the harm.

Corporate crimes are considered to be much more common and costly than people may think (Benson and Simpson, 2014). Illegal activities committed by corporation may result in catastrophic consequences in terms of human lives, economically and environmentally. Some scholars consider that corporate crime is perhaps one of the most dangerous crimes that occurs in our society (Simpson, 2002).

As Mokhiber (2007) clearly explains, corporate crime inflicts far more social and economic damage than all street crime combined. He claims that corporate crime is often violent crime, considering the astonish number of workers who die every year on their jobs or from occupational diseases, or the thousands of people 'who fall victim of the silent violence of pollution, contaminated food, hazardous consumer products and hospital malpractices'.

Countless social problems, such as recolonisation of developing countries, oppression of native and indigenous communities, food contamination, medical negligence and unsafe working conditions, are consequence of concentrated corporate power (Mokhiber and Weissman, 2005). What is more, many of the sanctions imposed on corporations have limited impact on their finance and business practices, hence the need for alternative deterrence approaches and effective control strategies of corporate crime.

In light of this, a deep analysis of a specific type of corporate crime will be proposed and implemented in the following chapters. In particular, all efforts will be focused on the understanding, modelling and analysis of accounting fraud offences. Consequently, an overview of the accounting fraud phenomenon is given next, along with a review of the most important accounting scandals of modern times as evidence of the social and economic impact of dishonest business reporting.

2.3 Overview of Accounting Fraud

The Association of Certified Fraud Examiners (ACFE) is one of the largest anti-fraud organisations responsible for providing anti-fraud training and education worldwide. In the ACFE's 2015 Fraud Examiners Manual, accounting fraud is defined as "the deliberate misrepresentation of the financial condition of an enterprise accomplished through the intentional misstatement or omission of amounts or disclosures in the financial statements to deceive financial statement users". Several synonyms of accounting fraud exist in the literature, including the so-called financial statement fraud, corporate fraud and management fraud.

According to Van Vlasselaer et al. (2015), there are five key characteristics that clearly distinguish between a genuine mistake and a fraudulent activity. They state that fraud is an "uncommon, well-considered, imperceptibly concealed, time-evolving and often carefully organised crime" that needs to be meticulously understood in order to achieve its early detection.

Furthermore, they say that by all means, fraud offences are not crimes that happen fortuitously but carefully planned instead. Fraudsters are not impulsive individuals; they are constantly adapting and perfecting their methods in order to maintain illicit moves undetected. The use of complex and organised schemes is very common when committing fraud, so the analysis should never be performed considering isolated events but rather the phenomenon as a whole.

It is also said that fraud is a corrupt behaviour, since it involves the misuse of entrusted power for personal gain (Baesens, Vlasselaer, and Verbeke, 2015). Accounting fraud is not directly associated with corruption by a politician or public servant, but by high-status senior executives using their privileged position for private benefits, regardless of whether they undermine the credibility of the organisation they promised to lead and manage, nor the dire consequences for their workers and customers.

2.4 Infamous Accounting Scandals

The dramatic increase of corporate accounting scandals in the last two decades accounts for the severity of this pandemic phenomenon, which often results in corporate bankruptcy, market collapse, economic crisis and more.

A brief explanation of some of the worst financial reporting scandals reported at large corporations is given next in order to illustrate the underestimated consequences of this unethical behaviour.

Enron, 2001

As mentioned briefly in Chapter 1.1, Enron Corporation, a giant energy company funded in 1985, engaged in a massive fraudulent scheme that culminated abruptly towards the end of the year 2001 with its impressive collapse and further bankruptcy. The main players of the manoeuvre, Chairman Jen Kay and CEO Jeffrey Skilling, kept billions of dollars in debt off the records, by using thousands of unconsolidated partnerships and very complicated financial reporting.

Eventually, figures did not match and the inevitable reduction in net income led to an approximate billion dollar-reduction in the equity of stockholders. Investors reacted immediately and soon Enron's stock price collapsed, precipitating the company towards an imminent bankruptcy. Consequently, Enron's shareholders lost nearly $74 billion and 4,500 employees lost their jobs and pension funds without proper notice (Swartz, 2003).

Even though the general opinion describes this massive collapse as unpredictable, Shilit and Perler (2010) affirm that the disaster could have been avoided if a careful examination of the public documents during the preceding years of the debacle had been performed. The impressive revenue growth from $9.2 billion in 1995 to $100.8 billion in 2000 should have warned the public, especially when considering that profits did not increase at such spectacular rate.

Enron's case is one of the most severe accounting scandals and audit failure of all times, considered by many analysts and experts as the largest corporate collapse in the U.S. history until Worldcom's bankruptcy the following year.

Worldcom, 2002

Less than a year after Enron's episode, the giant telecommunication services supplier Worldcom, now known as MCI, shook the financial market when one of the most serious *book-cooking* fraudulent schemes in U.S. history was uncovered, after an internal audit revealed important accounting errors from 1999 to 2002.

By means of aggressive accounting practices, CEO Bernie Ebbers and other major executives rigorously conceived a very complex plan that basically allowed the company to treat $3.8 billion of operating expenses as investment, moving regular expenditures from its statement of income to its balance sheets in order to exaggerate profits so as to create a false image of growth Shilit and Perler (2010).

The company filed for bankruptcy protection shortly after the exposure of the fraudulent scheme, leaving 17,000 employees laid off and an important loss of $180 billion assumed mostly by the investors and other external stakeholders.

Tyco, 2002

Tyco International Inc., a large security system company based on Princeton, New Jersey, achieved substantial growth mainly through the acquisitions of many large and small businesses. In fact, in a short 4-years period, more than 700 acquisitions were made, creating a fictitious image of growth that amazed their investors from 1999 to 2002.

The use of this particular methodology allowed Tyco to artificially increase earnings by means of creative accounting practices and accounting loopholes, directly hiding inconvenient transactions from the financial statements and improperly recognising exaggerated assets from the acquired companies.

Moreover, CEO Dennis Kozlowski and CFO Mark Swartz adopted a very uncommon and unethical praxis known as *comingling of assets*, that basically refers to the use of company's funds to pay for their personal expenses, such as properties, art pieces and private parties, amongst others. Both top ranking officers were charged later of stealing hundreds of millions of dollars from the company.

Parmalat, 2004

Italy's largest milk processor Parmalat engaged in misleading accounting practices consisting mainly of hiding gigantic debts associated with the expansionary strategy planned during the 80s and 90s.

In particular, Parmalat's debts were discovered to be eight times the figure the firm had admitted and reported. Moreover, a few months after the scandal came to light, Bank of America's former Chief of Corporate Finances in Italy admitted to have participated in a kickback scheme with executive and managers of Parmalat. It was also discovered, after a three-year trail, that several financial firms related to Parmalat's business and operations were participating in a complex market rigging strategy, including Bank of America, Citigroup, Deutsche Bank and UBS.

Investors and financial creditors filed a $10 billion class action suit against Parmalat's auditors and administrators, as well as Bank of America, Citigroup and external consultants, including Deloitte Touche and Grant Thornton.

Consequently, Tanzi family members, original owners of Parmalat, were arrested and sentenced along with several senior executives responsible for the fraud scheme. They were also investigated for intentional destruction of evidence and obstruction of justice.

American Insurance Group (AIG), 2005

The multinational insurance corporation AIG orchestrated a massive accounting fraud scheme that also involved a big-ridding and stock price manipulation.

Many accounting malpractices were carefully planned by CEO Hank Greenberg where huge loan figures were recorded as revenue, as well as secret arrangements were negotiated with fraudulent traders to further inflate stock prices.

In consequence, a $20 billion-dollar financial penalty was settled with the SEC. CEO Greenberg was fired but faced no criminal charges whatsoever.

Lehman Brothers, 2008

The collapse of Lehman Brothers culminated after they filed for bankruptcy in September, 2008. Lehman's bankruptcy filing was the largest in history considering that its assets exceeded those of previous bankrupt giants such as Enron and WorldCom, which contributed and intensified the 2008 financial crisis.

The scandal encompassed a colossal miscalculation of revenue and profit figures, which was at least doubtful considering the irregularities in the U.S. housing market and increment of defaults on subprime mortgages.

In particular, Lehman's high degree of leverage, calculated as the ratio of total assets to shareholders' equity, was enormous as a result of a portfolio of risky mortgage securities, situation that made the financial services firm increasingly vulnerable to the deteriorating market conditions.

Many measures were adopted by Lehman Brothers to regularise its business, but it was not enough. The stock price finally collapsed and the bankruptcy of the company was imminent.

Toshiba, 2015

Toshiba Corporation, a multinational conglomerate company headquartered in Tokyo, Japan, had been recently shaken by a massive accounting scandal connected to a $1.2 billion in overstated operating profits.

Accounting improprieties and book cooking practices were discovered over the course of seven years, strategy that was carefully planned by three former CEOs. Inappropriate accounting practices and overstated profits in multiple Toshiba business units began in 2008 along with the global financial crisis that had important repercussions in Toshiba's profitability.

The fraudulent scheme involved many tricks, including booking future profits early, pushing back losses, pushing back charges and other similar techniques that resulted in overstated profits.

2.5 Victims

Like any other criminal offence, accounting fraud is considered to be a social phenomenon since the "potential benefits for the fraudsters come at the expense of the victims" (Baesens, Vlasselaer, and Verbeke, 2015). Furthermore, it can be said that accounting fraud has important negative externalities to the society as a whole, that is, the cost to the community is much greater and harmful than the potential cost assumed by the company in case of detection.

This issue is critical because some of the resulting externalities can potentially be avoided by imposing stricter regulation in order to internalise them appropriately. Hence the importance of accurately evaluating the true costs of accounting fraud and all potential victims, including individuals, related businesses, government and the economy, among others.

Investors

Investors allocate their capital into an institution in order to receive a positive return in the future. There is always an intrinsic risk associated with an investment that is taken only if the expected gain remains attractive after considering the possibility of loss, thus the importance of making informed decisions regarding whether to invest in a particular company or not.

Consequently, investors should make an effort to investigate what are the healthier and more suitable options for them, and that requires putting faith in the investing world and presuming that the financial information released by the company is correct and truthful. Therefore, firms delivering erroneous financial reports end up betraying their investors since they are unexpectedly assuming greater risks that could lead to an important loss of capital.

For this reason, investors should maintain an active scepticism about financial reporting and also perform in-depth analysis of public documents to correctly evaluate businesses performance and position Shilit and Perler (2010).

Employees, Partners and Suppliers

As mentioned before, almost every accounting scandal results in massive collapse or bankruptcy, which involves large corporations to cease their businesses and consequently, hundreds of workers losing their jobs and sometimes also their pensions.

In addition, when a large company goes bankrupt, an important drop of its demand occurs. In consequence, many related businesses, such as partners and suppliers, are left without clients and subsequently out of business.

Government and Market System

Accounting scandals are usually related to audit failure, not just because internal auditors missed or ignored falsified reports but also because public examination failed to detect them in an early stage. This typically occurs when a fraud opportunity appears, which means securities laws, rules and regulations were not sufficient to ensure the proper operation of the economic system. Consequently, government authorities, such as regulatory agencies and taxation offices, suffer a significant loss of credibility, as well as the market system.

A very sensitive issue is faced here because one of the most precious pillars of the economy is violated, that is, trust. Thus, in order to restore the economic confidence, the introduction of amended initiatives usually takes place as public policy measures seeking to improve ethical financial reporting and to promote better accounting practices. Such is the case of the renowned Sarbanes-Oxley Act, enacted on July 30, 2003, as an attempt to make corporate accounting more transparent and further protect investors and the securities market.

Creditors and Financial Analysts

Both creditors and financial analysts have to make daily financial decisions on whether to extend credits or invest in different companies, therefore they must rigorously analyse their financial statements and business positions, among other attributes.

That is to say, fraudulent reporting will most probably contribute to erroneous creditworthiness assessments and investment decisions, due to the imprecise evaluation of associated risks; again, undermining market confidence and debilitating the proper operation of the economic system.

All the above clearly expose the negative social and financial impact of accounting fraud, and also evidence the urgent need for government and regulatory agencies to invest in accurate fraud-detection systems.

2.6 Forensic Accounting

The main goal of fraud detection is to discover hidden patterns of fraudulent activities in order to expose them as soon as possible and, therefore, rapidly address recovery strategies and attenuate potential losses.

Accounting fraud is a complex and dynamic phenomenon, hence the need for a deep understanding of the underlying behaviour when designing a detection mechanism. As such, an expert-based approach has to be built that takes into account previously detected fraud cases to further establish rules and indicators to be used to discover new cases of fraud (Baesens, Vlasselaer, and Verbeke, 2015).

Current detection methods typically require manual investigation of suspicious cases that heavily relies on human expertise, which is an incredibly costly task in terms of time and labour. Thus, the usefulness of statistical models to help identify questionable accounts and to further define guidelines for detecting the occurrence of a fraudulent activity is fairly clear. Most notably, Bolton and Hand (2002) claim that fraud detection is an important area where "statisticians can make a very substantial and important contribution", since existing methods are outdated and, in consequence, perform poorly examination of public documents.

Broadly speaking, a detection mechanism should satisfy three essential conditions in order to be successful, including accuracy, interpretability and efficiency (Baesens, Vlasselaer, and Verbeke, 2015). Accuracy, in terms of achieving high levels of detection power, specifically when dealing with suspicious cases, as well as the ability to generalise results to unknown observations; interpretability, as to properly communicate the relevant information in a simple manner so it can be integrated and later utilised by all interested parties; and efficiency, with regard to meeting time constraints, operational requirements and cost restrictions.

All of the above clearly shows the convenience of a data-based approach to help conceive an adequate and integral fraud-detection method, to be used to identify clear signs of fraudulent financial reporting and further assist a more comprehensive and objective examination of corporate financial reports.

Lastly, it is worth mentioning that several forensic accounting-similar approaches have been adopted in related fields to, for instance, identify money laundering (Ravenda, Argilés-Bosch, and Valencia-Silva, 2015) and organised crime infiltration in organisations and firms (Savona and Berlusconi, 2015), which evidence the usefulness of this kind of methodologies to detect, predict and tackle different corporate and financial crimes.

2.7 Literature Review

Prior Studies Overview

Part of the fraudulent financial reporting literature has focused primarily in the evaluation of qualitative characteristics related to the board of directors and principal executives, including information of corporate governance structure (Beasley, 1996; Hansen et al., 1996; Bell and Carcello, 2000) and insider trading data (Summers and Sweeney, 1998). Studies using this kind of information show promising results; however, getting access to such data is very difficult and sometimes even prohibited for most individuals.

On the other hand, studies using publicly available financial statement information are less common and usually incorporate small samples. Generally, the selection of fraud cases is limited to certain conditions and manually matched with non-fraud observations on the basis of business fundamentals, such as industry, size, maturity, period and more. Undoubtedly, there is an interesting gap in this area of the literature where the selection process of a more representative sample has the potential to be explored and expanded, which is intended to be filled in the present study.

With regard to the employed techniques, discriminant analysis and logistic regression are by far the most popular. Such algorithms are commonly considered as a benchmark framework due to their simplicity and low computational cost, and because they have been proven to efficiently detect falsified accounting reporting in relatively small samples (Fanning and Cogger, 1998; Spathis, Doumpos, and Zopounidis, 2002; Kaminski, Wetzel, and Guan, 2004; Pai, Hsu, and Wang, 2011).

In particular, discriminant analysis (DA) has been used in a variety of disciplines, including bankruptcy prediction of public companies (Altman, 1968), marketing research (Crask and Perreault, 1977) and medical studies (Yarnold, Soltysik, and Martin, 1994; Yarnold et al., 1995), amongst others. However, results of the application of DA for detecting accounting fraud are not very promising, as can be seen in Kaminski, Wetzel, and Guan (2004), where an exploratory analysis is designed to investigate whether financial ratios are useful to detect fraud, concluding that there is no significant difference in the ratios of fraudulent versus non-fraudulent firms. The results obtained in this study are fairly inaccurate, since the absolutely opposite is shown in what remains of this thesis. In fact, it has been found that financial ratios are significant predictors of accounting fraud offences, as important differences can be identified when comparing fraudulent and non-fraudulent corporations.

A great deal of work has been done using logistic regression (LR), as an alternative approach to discriminant analysis. For instance, Persons (1995) applies two stepwise-logistic regression models, one to predict fraud in the first year of occurrence and the other for the preceding year. Results suggest that both models outperform a naive strategy of classifying all firms as non-fraud and that it is easier to detect firms that are most likely to commit fraud (using preceding-year information), than to detect fraudulent firms (using fraud-year information).

Likewise, Lee, Ingram, and Howard (1999) develop a logistic regression model to examine the relationship between earnings over operating cash flow and accounting fraud. The results indicate that the excess of the difference between the aforementioned financial items is extreme in most fraud firms in years immediately prior to the fraud violation. Spathis (2002) also applies a logistic procedure, demonstrating that the model performs effectively in detecting fraudulent reports, since it correctly classifies 84% of all cases. Initially, he uses a set of 17 financial ratios as explanatory variables, and concludes that only ten are selected as potential predictors of false financial reporting. He concludes that variable selection can be very useful for accounting fraud detection, claim that is supported by the results obtained in the present study.

In a similar fashion, Lenard, Watkins, and Alam (2007) use a logistic regression approach to predict falsified financial statements in service-based computer and technology firms. In addition to popular financial ratios, they include a 'fuzzy logic' variable to assess the impact of non-financial red-flags in the occurrence of fraudulent reporting. The proposed model shows an overall accuracy of 77%, suggesting that the inclusion of the proposed variable enhanced the classification accuracy significantly.

More recently, Dalnial et al. (2014) analyse the usefulness of financial ratios as predictors of fraudulent accounting reporting. They use a sample of Malaysian public firms for training a logistic regression model, reaching an overall accuracy of 75% and concluding that several financial ratios such as total debt to total assets and receivables to revenue are significantly helpful for accounting fraud detection.

Decision trees (DT) are another well-known machine learning method often used to predict fraudulent accounting records, mainly due to their fewer data preparation requirements and their intuitive interpretation. To illustrate, Gupta and Gill (2012) design a data mining methodology for preventing and detecting accounting fraud, concluding that decision trees are superior in terms of accuracy than other statistical methods such as genetic programming, since they correctly classified 95% of all cases, specifically 98% of non-fraudulent firms and 86% of fraudulent firms.

Similarly, Pai, Hsu, and Wang (2011) introduce a fraud warning model to assess the likelihood of falsified financial reports. They build a support vector machine for detecting fraudulent firms and then develop a decision tree model, in order to establish easy-to-grasp rules that can be used by auditors for accounting fraud detection. The proposed algorithm outperforms the other three approaches - discriminant analysis, logistic regression and neural networks - in terms of testing accuracy.

Alternative approaches have also been adopted in order to detect accounting fraud: Neural networks (NNs) are particularly popular for this end, showing promising results when predicting fraudulent accounting practices. Kwon and Feroz (1996) investigate the efficacy of several red flags in predicting reporting violations. They compare the results of a neural network and a logistic regression model, concluding that the neural network approach outperformed by more than 40% in terms of average classification accuracy.

In addition, Choi and Green (1997) develop three back-propagation NNs using different expectation methods in order to transform the percentage change between the reported and the expected account balance of the fraud-year. The results are robust for all the three models, since Type I and Type II errors are significantly less than a naive strategy of random choice, suggesting that neural networks have significant potential as accounting fraud detection tool.

Furthermore, Fanning and Cogger (1998) conclude that the use of public documents is particularly helpful for detecting falsified financial statements and that neural networks greatly outperform standard statistical methods for predicting fraudulent reporting practices, such as logistic regression and discriminant analysis. In addition, Feroz et al. (2000) test the ability of NNs and conclude that they perform better than conventional logistic regression, and confirm that financial ratios calculated from publicly available data have significant predictive value.

Later on, Kirkos, Spathis, and Manolopoulos (2007) compare the relative performance of NNs and other techniques, such as decision trees and Bayesian networks, using a stratified 10-fold cross validation approach. They concluded that Bayesian networks generally outperform the other two methods, while decision trees exhibit the lowest performance in terms of classification accuracy.

A last application of a NNs approach can be seen in Ravisankar et al. (2011) where they examine the usefulness of various data mining techniques - neural networks, support vector machines, genetic programming and logistic regression - for the detection of fraudulent Chinese firms, with and without feature selection. Results suggest that the t-statistic technique is a simple and effective approach for selecting significant features and that NNs outperformed all remaining methods in both scenarios.

More complex settings have been proposed, using an assemblage of several machine learning methods to better understand the accounting fraud phenomenon and to improve its detection rate. For example, Kotsiantis et al. (2006) develop a hybrid decision support system that merges different techniques such as decision trees, artificial neural networks, Bayesian networks, logistic regression and support vector machines, achieving better results compared to the aforementioned methods employed individually.

Later on, Song et al. (2014) propose an ensemble of four machine learning techniques, including logistic regression, decision trees, neural networks and support vector machines, in order to assess the risk of fraud. The experimental results indicate that the suggested approach outperform the above methods showing a classification accuracy of 89%.

Nevertheless, the achieved performance of neural networks and more complex methodologies, such as Bayesian networks, support vector machines and hybrid algorithms, is counteracted by the considerable drawbacks that these methods entail, including important computational costs and overfitting proneness, as well as struggling when interpreting results (Tu, 1996).

Finally, Shilit and Perler (2010) conceive a guide to detect accounting gimmicks and fraud in financial statement reports. They discover several financial shenanigans adopted by fraudulent firms, including improper recording of revenue, shifting of expenses and/or income to other periods and irregular disclosure of liabilities, among others. Consequently, the authors conclude that many accounting scandals and corporate collapses could have been avoided if a careful examination of the public documents during the preceding years of the event had been performed, and that using the suggested clues could be beneficial to further warn the public before a disaster occurs.

A summary table including all prior studies using machine learning techniques is provided next. More details can also be observed (Table 2.1), such as sample size, number of fraud cases, methods employed and overall accuracy when available.

Predictors of Accounting Fraud

Different and diverse explanatory variables have been considered in previously mentioned studies, including financial statement items, financial ratios and corporate governance information. Nevertheless, the most common predictors adopted for accounting fraud detection are financial ratios, mainly because it has been shown that these types of variables have a great predictive capacity.

Table 2.2 summarises the most popular financial ratios used in the literature and the reference to the studies at issue. A more detailed discussion of these and more financial ratios is performed in Chapter 4, as well as a comprehensive analysis of how they are manipulated, the relationship between them and their statistical significance when detecting accounting fraud.

Research Considerations

Many contributions can be attributed to prior studies, as all accounting fraud research enhance awareness and knowledge of this phenomenon and further support its detection and anti-fraud preventive measures. However, a great deal of work that can be further done to improve detection strategies in many ways, including sample size, industries at issue, machine learning methods and evaluation metrics.

Firstly, it can be observed from Table 2.1 that sample sizes of previous studies are fairly small. In most studies, samples are manually selected, which is a highly problematic practice as it is inherently biased and so results cannot be extrapolated to the population. Therefore, increasing the amount of data used to train and test the models is a noticeable enhancement, as well as attempting to collect as many fraudulent cases as possible, and not only the most convenient for the sake of research results.

Secondly, most prior studies focus their analysis in specific industries defined by the Standard Industrial Classification (SIC) system. After careful review, it is surprisingly

TABLE 2.1: Prior studies in detecting accounting fraud

Study	Sample Size	Fraud Cases	Method(s)	Overall Accuracy (%)
Persons (1995)	206	103	Logistic Regression	n/a
Kwon & Feroz (1996)	70	35	Neural Networks Logistic Regression	88 47
Choi and Green (1997)	172	86	Neural Networks	n/a
Fanning & Cogger (1998)	204	102	Logistic Regression Discriminant Analysis Neural Networks	50 52 63
Lee et al. (1999)	620	56	Logistic Regression	n/a
Feroz et al. (2000)	132	42	Neural Networks Logistic Regression	81 70
Spathis (2002)	76	38	Logistic Regression	84
Spathis et al. (2002)	76	38	Multicriteria Decision Aid Method Discriminant Analysis Logistic Regression	88 84 81
Lin et al. (2003)	200	40	Neural Networks Logistic Regression	76 79
Kaminski et al. (2004)	158	79	Discriminant Analysis	n/a
Kotsiantis et al. (2006)	164	41	Decision Trees Neural Networks Bayesian Networks Logistic Regression Support Vector Machines Hybrid Decision Support System	91 80 74 75 79 95
Kirkos et al. (2007)	76	38	Decision Trees Neural Networks Bayesian Networks	74 80 90
Hoogs et al. (2007)	390	51	Genetic Programming	n/a
Lenard et al. (2007)	30	15	Logistic Regression	77
Ravisankar et al. (2011)	202	101	Support Vector Machines Genetic Programming Logistic Regression Neural Networks	72 89 71 91
Pai et al. (2011)	75	25	Support Vector Machines Discriminant Analysis Logistic Regression Decision Trees Neural Networks	92 81 79 84 83
Gupta & Singh (2012)	114	29	Decision Trees Genetic Programming	95 88
Danial et al. (2014)	130	65	Logistic Regression	75
Song et al. (2014)	550	110	Logistic Regression Decision Trees Neural Networks Support Vector Machines	78 79 85 86

TABLE 2.2: Most popular explanatory variables for detecting AF

Predictor	Study
NITA	Persons (1995) Spathis (2002) Spathis et al. (2002) Kaminski et al. (2004) Kirkos et al. (2007) Lenard et al. (2007) Ravisankar et al. (2011) Pai et al. (2011) Gupta et al. (2012) Danial et al. (2014) Song et al. (2014)
TLTA	Persons (1995) Spathis (2002) Spathis et al. (2002) Kaminski et al. (2004) Kotsiantis et al. (2006) Lenard et al. (2007) Pai et al. (2011) Song et al. (2014)
WCTA	Spathis (2002) Spathis et al. (2002) Kaminski et al. (2004) Kotsiantis et al. (2006) Kirkos et al. (2007) Pai et al (2011)
RVSA	Fanning & Cogger (1998) Feroz et al. (2000) Kaminski et al. (2004) Pai et al. (2011) Schilit and Perler (2010)
SATA	Fanning & Cogger (1998) Spathis (2002) Kotsiantis et al. (2006) Kirkos et al. (2007) Lenard et al. (2007)
CACL	Kotsiantis et al. (2006) Lenard et al. (2007) Ravisankar et al. (2011) Song et al. (2014)
IVSA	Fanning & Cogger (1998) Spathis (2002) Spathis et al. (2002) Pai et al. (2011)
IVTA	Ravisankar et al. (2011) Gupta et al. (2012) Danial et al. (2014) Song et al. (2014)

observed (Table 2.1) that there are no studies that investigate accounting fraud within financial services firms. The main reason for this exclusion is that they are structurally different and an alternative set of variables may be required since certain financial statement items, such as accounts receivable and inventory, are not available for these companies. Hence "research to find the variables most useful in the specific industries would be of great value", especially in the poorly examined area of financial services (Fanning and Cogger, 1998). As such, a substantial improvement is achieved in the present study as cases from all industries are included.

In brief, it can be said that although the existing techniques have increased the detection rate of accounting fraud offences, these are very limited and often not sufficient to uncover complex fraudulent schemes, hence the need for improved methodologies that comply with the aforementioned basic principles of accuracy, interpretability and efficiency (Chapter 2.6).

2.8 Proposed Methodology

By means of this thesis, it is proposed to implement an analytical approach that pursues the detection and control of accounting fraud offences. For this reason, an Accounting Fraud Detection and Control methodology will be elaborated that involves four key elements: (i) the accounting element as a financial concept; (ii) the fraud element determined by a particular criminal behaviour; (iii) the element of detection associated with the exposure of the criminal behaviour; and (iv) the element of control that implies more effective detection strategies.

The first element of accounting is directly associated with how the economic system works. Most modern corporations are always looking to maximise their profits, which are more certainly monetary. As such, accounting reporting is the best way for companies to communicate to internal and external stakeholders about the operation of their businesses. Consequently, information about corporate performance, activities and decisions are usually summarised in financial statements and reports, which are expected to be align with universal generally accepted accounting standards. In light of this, it is reasonable to suggest close examination of abnormal behaviour and suspicious patterns that may be indicating illegitimate corporate activities.

For this end, it is required to understand the phenomenon and to identify most common fraudulent practices. In other word, to recognise fraudulent criminal behaviours (second element). The identification of criminal behaviours and abnormal financial patterns should be accomplished objectively and accurately, which is why statistical modelling is so attractive in this context. A representative modelling of the accounting fraud phenomenon will greatly support the effective exposure of the criminal behaviour, which relates to the third element of the proposed approach.

In order to improve the detection of accounting fraud offences, analytical models are required first to identify companies that behave in an abnormal way, and second to recognise what accounting tricks are these firms employing to hide poor financial performance. Once the modelling task is accomplished, guidelines for more effective and efficient examination of financial reports are fairly easy to obtain and further be included as part of a comprehensive and adaptive control strategy, which correspond to the final element of the suggested methodology.

Considering all of the above, it is proposed to apply a forensic data analysis approach to: (i) create an extensive accounting fraud database consisting of fraud cases from all industry areas and financial statement information for both fraud and non-fraud firms; (ii) handle the collected data set; (iii) evaluate differences in financial accounts between corrupted and genuine reports; (iv) examine distinctive characteristics of fraudulent and non-fraudulent financial statements; (v) implement several machine learning methods in order to better differentiate between fraud and non-fraud cases; and ultimately (vi) identify industry-specific financial indicators to be used as 'red flags' for accounting reporting examination. These are, thereby, the specific goals of this thesis.

2.9 Summary

In order to properly approach the phenomenon of accounting fraud, a comprehensive understanding of corporate financial malpractices was required. In this regard, an overview of the wider topics of white-collar crime and corporate crime is elaborated, followed by an exhaustive analysis of the behavioural aspect of accounting fraud, most severe cases of accounting scandals and potential victims.

In addition, a critical literature review has been performed to account for research achievements and good practices, as well as to identify relevant shortcomings related to accounting fraud analysis and further detection. Finally, taking everything into consideration, a thorough methodology has been proposed to ultimately achieve the desired research objectives.

Chapter 3

Forensic Data Analysis

3.1 Forensic Analytics

Accounting fraud perpetrators are continuously conceiving new ways to commit their offences and, in consequence, always transforming their fraudulent behaviour, thus the complexity of the accounting fraud phenomenon. This deliberate managerial wrongdoing is particularly hard to detect and predict, since it involves deep knowledge of accounting and legal tricks that are intentionally employed to make documents look genuine and error-free. As such, a data-driven detection mechanism is suggested, based on publicly available financial statement information, to be used by all related parties interested in discovering sophisticated fraudulent schemes.

Forensic data analysis is concerned with the treatment and examination of financial crime offences, hence the relevance of its use to develop an adequate technique for fraud detection. In particular, a forensic accounting approach is proposed in order to overcome potential auditing failure and further improve examination of public documents through the recommendation of meaningful examination of accounting items.

One of the most important elements of a data-driven mechanism is, without a doubt, the data. Not only the tangible record of fraudulent and non-fraudulent offences, but also how it is registered and displayed, what transformation and further handling procedures are needed, what additional information should be included, which should be removed and what sample should be selected.

It is generally known that valuable analytical models rely on clean and organised data to be used to generate the required methodological conditions before proceeding with further analysis (Baesens, Vlasselaer, and Verbeke, 2015). Accordingly, data collection, preparation and validation play a crucial role in forensic analytics, particularly when dealing with financial statement items as they are expected to align with universal generally accepted accounting standards.

A careful description of the implemented data handling process and sample selection methodology is explained in what follows.

3.2 Fraud Data Collection

The data collection task is critical in crime-related research, since it is very difficult to find sufficient and accurate data for analysis. In addition, and given the highly sensitive nature of the topic, there is a limited amount of relevant journal articles related to accounting fraud detection, and publication of controversial results may be censored or even prohibited (Bolton and Hand, 2002). Therefore, a compilation of an exhaustive and representative database containing relevant cases of accounting fraud instances is imperative to further achieve the proposed objective.

In this study, accounting fraud cases are identified considering all Accounting Series Releases (ASR) and Accounting and Auditing Enforcement Releases (AAER) issued by the U.S. Securities and Exchange Commission (SEC) between 1990 and 2012. Particularly, all public litigation releases involving deceptive reporting were hand-collected first from the SEC's website[1] and then cross-validated with an official fraud-database provided by the Securities and Class Action Clearinghouse (SCAC), Stanford Law School. This accredited data set was obtained after the enactment of a non-disclosure agreement between the involved parties.

The selection of the studied period is justified based on data availability and practicality considerations. On the one hand, discovered fraud cases published by the SEC include successful enforcement actions with monetary sanctions exceeding $1 million announced between July 29, 2002 and present. Accounting fraud cases released by the SEC date from 1990 onwards, hence the selection of the year 1990 as the beginning of the studied period. On the other hand, this study began in the middle of 2013, so including this year would be incorrect considering that many cases

[1] SEC Sanctions Database: https://www.secwhistlebloweradvocate.com/program/sec-enforcement/sanctions-database/

of fraud could be discovered in the remainder of the year. As such, 2012 is selected as the final year of the studied period.

Lastly, two main considerations should be taken into account regarding the collected data. First, non-public firms were excluded as the SEC only has jurisdiction over publicly traded companies. Second, the non-fraud cases collected may contain firms that have engaged in fraud but have not been discovered, which could be influencing findings and results. More about the latter is discussed in Chapter 8.

3.3 Financial Data Collection

In addition to the aforementioned compilation of publicly known fraud offences, a collection of financial statement information is required for further modelling, as this data will be used to identify frequent accounting tricks adopted to improperly modify financial reports. The rationale behind the use of financial statement data is that this source of information should be enough to fairly reveal the value of a firm (Ou and Penman, 1989). A relevant analysis of financial statements generally allows the extraction of significant information regarding the true financial worth of an institution, its accounting structure and financial performance.

Accordingly, the collection of published financial statement data corresponding to all public companies was gathered from the COMPUSTAT files in the interest of creating an integrated database containing relevant accounting information related to fraudulent and non-fraudulent firms for the pertinent time period, that is, from 1990 to 2012.

COMPUSTAT is a data-repository containing several databases related to financial, statistical and market information from companies all around the world, covering 99% of the world's total market capitalisation. Many interested parties benefit from the information provided by this platform, such as investors, academic researchers, bankers, analysts, advisors and portfolio managers, among others. COMPUSTAT provides a broad range of relevant information, including annual and quarterly business fundamentals, and pricing and property data. More than 300 annual and 100 quarterly documents can be found, covering information related to Income Statement, Balance Sheet and Cash Flows, as well as supplemental data items on more than 24,000 active and inactive publicly held companies.

Financial statements are formal documents reflecting the financial status of a company. It is required that all publicly traded businesses release this information every quarter and year so it can be audited by government agencies and public auditors, amongst others, in order to ensure its legitimacy and accuracy. The data provided in the financial statements usually involve records associated with income statements, balance sheets and cash flows, and as such, will be considered as relevant financial information to be used for statistical modelling.

In particular, 17 financial items are collected from annual financial statements since they are required to further create explanatory variables commonly used as fraud predictors by previous studies (Chapter 2.7). A brief explanation of the chosen financial items is given using the official definition provided by COMPUSTAT, unless otherwise specified.

3.3.1 Balance Sheet

The balance sheet statement is one of the key sources of data for analysing the book value of a company. It is usually published at the end of the fiscal year and comprises three main business components: assets, liabilities and shareholders' equity. The relationship of these items is expressed in the fundamental balance sheet equation given by the following formula:

$$Assets = Liabilities + Equity \tag{3.1}$$

In other words, it can be said that the net worth of the company is the difference between its assets and liabilities. A detailed description of these three elements can be found next:

1. **Assets**: all tangible and intangible resources with economic value that a company owns and has in its possession, or that eventually will receive and acquire as its property.

   Selected financial items related to assets are defined below:

   - Total Assets (TA): This item represents the total assets of a company at a point in time. If the company does not report a useable amount, this data item will be left blank.

- Current Assets (CA): This item is a component of Total Assets (TA) and represents cash and other assets that are expected to be realised in cash or used in the production of revenue within the next 12 months. This item is not available for banks.
- Cash (CH): This item is a component of Cash and Short-Term Investments which in turn is a component of Current Assets (CA), and represents any immediately negotiable medium of exchange or any instruments normally accepted by banks for deposit and immediate credit to a customer's account.
- Accounts Receivable (RV): This item is a component of Current Assets (CA) and represents asset designation applicable to all debts, unsettled transactions or other monetary obligations owed to a company by its debtors or customers[2].
- Inventory (IV): This is a component of Current Assets (CA) and represents merchandise bought for resale and materials and supplies purchased for use in production of revenue.

2. **Liabilities**: all financial debt or obligations that must be paid under contractual conditions and time frames. Liabilities are typically used to finance operations and pay for potential business expansions.

   Selected financial items related to liabilities are defined below:

   - Total Liabilities (TL): This item represents current liabilities plus long-term debt plus other non-current liabilities, including deferred taxes and investment tax credit.
   - Current Liabilities (CL): This item is a component of Total Liabilities (TL) and represents liabilities due within one year, including the current portion of long-term debt. This item is not available for banks.
   - Accounts Payable (PY): This item is a component of Current Liabilities (CL) and represents only trade obligations due within one year or the normal operating cycle of the company.

[2] Definition obtained from Investopedia: http://www.investopedia.com/terms/r/receivables.asp

- Long-Term Debt (LTD): This item is a component of Total Liabilities (TL) and represents debt obligations due more than one year from the company's balance sheet date.

3. **Shareholders' Equity**: retained earnings and funds contributed by shareholders that take the risk of investing in a particular company. It can be seen as the return on stockholders' investment or, in other words, shareholders' ownership of the company's assets.

   Selected financial items related to equity are defined below:

   - Total Equity (TE): This item represents the common and preferred shareholders' interest in the company.
   - Retained Earnings (RE): This item is a component of Total Equity (TE) and represents the cumulative earnings of the company less total dividend distributions to shareholders.

One last financial item related to the balance sheet statement not contained in the categories of Assets, Liabilities or Equity, is defined below:

- Working Capital (WC): This item represents the difference between total current assets minus total current liabilities as reported on a company's Balance Sheet. This item is not available for banks.

3.3.2 Income Statement

The Income Statement is a financial report that accounts for a company's earnings for a given time period. In particular, it shows the incoming revenues for the specified period in addition to the associated outgoing expenses.

Selected financial items related to income statements are defined below:

- Net Income (NI): This item represents the fiscal period income or loss reported by a company after subtracting expenses and losses from all revenues and gains.
- Total Sales (SA): This item represents gross sales (the amount of actual billings to customers for regular sales completed during the period) reduced by cash discounts, trade discounts, and returned sales and allowances for which credit is given to customers, for each operating segment.

- Cost of Good Sold (COGS): This item represents all costs directly allocated by the company to production, such as material, labour and overhead.
- Earnings Before Interest and Tax (EBIT): This item is the sum of Total Sales (SA) minus expenses, excluding tax and interest[3].

3.3.3 Cash Flow

The Cash Flow statement is a financial report that describes incoming and outgoing funds in both balance sheet and income statements that affect cash and cash equivalents. In other words, it specifies the movement of cash in and out of a business. In particular, it shows all flows related to operating, investing and financing activities.

Selected financial items related to cash flow statements are defined below:

- Cash Flow From Operations (CFFO): This item represents the net change in cash from all items classified in the Operating Activities section on a Statement of Cash Flows, where increases in cash are presented as positive numbers and decreases in cash appear as negative numbers. This item is not available for banks.

3.4 Data Preparation

A rigorous preparation of the collected data must be conducted after the first stage of collection. As explained, two sources of information have been collected, including accounting fraud cases and financial statement data, both in need for profound handling and preparation, as raw data being noticeably messy.

Data preparation is specially challenging considering the size of the resulting dataset in terms of number of observations and number of variables. Preparing the data includes mainly the tasks of data cleaning, data transformation, data merging, treatment of missing values and data validation, all carefully described and documented below.

[3] Definition obtained from Investopedia: http://www.investopedia.com/terms/e/ebit.asp

3.4.1 Data Cleaning

In the context of accounting fraud, the data cleaning task primarily involves ensuring that the sample collected includes nothing but recognisable public companies within the period of interest. Consequently, several exclusions were required and further conducted, such as:

- Removal of all fraud and non-fraud instances that occurred before the year 1990 and after the year 2012.
- Removal of duplicated data. Duplications occur when two or more observations share the same information regarding the company and the year at issue. Companies are easily identifiable using features such as *company name*, *GVKEY*, *ticker symbol*, *CUSIP* and *CIK number*. Years at issue are based on fiscal year indexing rather than calendar data.
- Removal of firms with undisclosed or non-applicable/unclear identification, again based on identifying features as mention above.

3.4.2 Data Transformation

All collected instances related to fraud offences are fully defined by the company that committed the violation, as well as the period of time in which the violation occurred. In this way, when a fraud is perpetrated for more than a year, then it is required to partition the case into year-instances to further combine them with the relevant annual financial statement information.

This procedure results in 1,594 fraud-year observations again characterised entirely by company I.D. and the associated fiscal year of the offence. Table 3.1 summarises the number of fraudulent observations obtained after splitting fraud cases into the corresponding years of occurrence, particularly arranged by industry area of where companies belong to.

More details on the subsectors included in each industry will be extensively discuss in Chapter 4.

TABLE 3.1: Fraud cases by industry

SIC Codes	Standard Industrial Classification (SIC)	Fraud Cases	Perc (%)
0100 - 0999	Agriculture, Forestry and Fishing	11	0.69
1000 - 1799	Mining and Construction	52	3.26
2000 - 3999	Manufacturing	609	38.21
4000 - 4999	Transportation, Communications, Electric and Gas	106	6.65
5000 - 5999	Wholesale Trade and Retail Trade	169	10.60
6000 - 6799	Finance, Insurance and Real Estate	236	14.81
7000 - 8999	Services	375	23.53
9100 - 9729	Public Administration	36	2.26
		1,594	**100**

3.4.3 Data Merging

In order to apply a forensic analytical approach, it is convenient to present the collected information in a single table as it is easier to process and analyse the data when it is stored in a structured manner. Occasionally, when data tables are combined, important errors are made especially when a great deal of observations are involved. Hence the need for detailed scrutiny of the resulting tables to make sure that the data is correctly integrated (Baesens, Vlasselaer, and Verbeke, 2015).

On that account, both collected sources of information, that is, fraud cases by year and annual financial statement data, have to be merged. These datasets are combined into one final table using company I.D. and year as merge keys. The resulting database consists of all public companies existing during the studied period, their identifying information, company description, financial statement items and the corresponding fraud or non-fraud flag.

3.4.4 Data Validation

Once a comprehensive data table has been conceived, the validation of the collected data must be performed, as only valid instances should be considered for further analysis. Nevertheless, the validation of the data is a different process depending on the context in which the study is being conducted.

In the case of accounting data, financial statement validation is a very important part of the data preparation process. A common practice when validating financial statements is what is called *accounting reconciliation*, which basically consists in ensuring that two or more figures are accurate and in agreement. This process is mainly used to determine whether financial amounts match across different sections of financial reports, and that calculations are made in a legitimate manner.

That being said, the following account reconciliations are evaluated and justified when appropriate:

- Ensure that equation 3.1 is met, that is, the sum of the values related to total liabilities and shareholders' equity is equal to total assets.
- Ensure total assets are positive figures, since it is inaccurate to define assets as negative figures, as well as businesses that exist without total assets.
- Ensure total liabilities are not negative, as negative values are most likely invalid observations.
- Ensure working capital is equal to the difference between current assets and current liabilities, as properly defined in Section 3.1.
- Ensure total sales figures are positive, otherwise there is no economic activity, which certainly suggests invalid observations.
- Ensure inventory figures are not negative, as negative values have no economic meaning.
- Ensure accounts receivable are not negative, as negative figures have no economic meaning either.

3.4.5 Missing Values

Missing values are very common in the process of data preparation, especially when working with COMPUSTAT data. This situation usually occurs due to differences in reporting formats or structural changes of the databases over time. On the one hand, reports may differ from firms belonging to different industries or that have presence in foreign countries, hence dissimilar accounting standards will determine their financial accounts. On the other hand, it is also possible that COMPUSTAT databases are adjusted from time to time, resulting for example, in new variables that only contain entries for current periods and not for older years, or conventional variables that become meaningless when replaced by up-to-date features.

Several techniques are typically employed to deal with missing data, including: (i) keeping missing values as a separate category, particularly when considered to be meaningful; (ii) replacing missing values with the value of zero; (iii) replacing the missing value with a well-known central tendency measure, such as mean, median or mode; (iv) replacing missing values with an estimated figure obtained from a regression-based technique; or simply (v) deleting observations or variables with a great number of missing values (Baesens, Vlasselaer, and Verbeke, 2015).

The most straightforward option of deleting observations with missing values, is adopted in the present study. The rationale behind this decision is that speculating on the legitimate value of a particular financial item for a particular entity seems fairly irresponsible considering how sensitive the topic of accounting fraud is. Estimated financial figures may be reasonably close to the actual value, but when not the case, fictitious results may be obtained from analytical models, and in consequence, incorrect classification of non-fraudulent and fraudulent firms. This is an interesting exercise that could be tested in future work.

Thereby, observations containing missing values within relevant financial items, will be deleted as follows:

- Drop observation if total assets figure is missing: undisclosed values of total assets have no practical meaning as no firms can exist without assets.
- Drop observation if missing sales: missing total sales figures represent non-existent economic activity, hence the business is not valid.
- Drop observation if missing net income: similarly, no income means no economic activity, thus not a valid observation.

As a result of the adopted missing values treatment methodology, 15% of the database originally collected has been dropped.

3.5 Sample Selection

One of the main characteristics that defines the fraud phenomenon so uniquely is that it is an uncommon activity (Chapter 2), particularly in the context of accounting fraud since only a minority of the recorded cases are actually classified as fraudulent.

Learning from these rare events is a very challenging task given the small amount of observations available to train predictive models, hence especially difficult to further discriminate between fraudulent and non-fraudulent instances. As Cerullo and Cerullo, 1999 express in regards to this matter, “unrepresentative sample data or too few data observations will result in a model that poorly estimates or predicts future values".

The class-imbalance problem fully emerges when statistical learning models are applied, because they all opt for a naive strategy of classifying all firms as non-fraudulent. As a consequence, accuracy measures show excellent average performance that only reflect the underlying uneven class distribution. Nevertheless, the methods are totally ineffective in detecting positive cases (Chawla, Japkowicz, and Kotcz, 2004).

Therefore, the selection of a representative sample is required in order to solve the imbalance problem encountered in this study, and also to enhance the discriminatory power of the proposed statistical models. The number of fraud cases in the collected dataset only represents a 0.8% of all observations, hence a stratified sampling method is implemented for the selection of non-fraud cases.

Thereby, the stratifying exercise is conducted according to the target variable *Fraud*, as it matches exactly the same amount of fraud observations as in the original data, specifically on the basis of industry area and fiscal period. Consequently, the sample selection process occurs in two phases, first dividing the dataset by industry and then by year.

A variety of sampling methods can be employed when dealing with imbalanced datasets, individually or in combination, hence an extensive and interesting analysis could be done to select suitable samples of fraudulent and non-fraudulent cases. A more detailed discussion about this topic is addressed in Chapter 8.

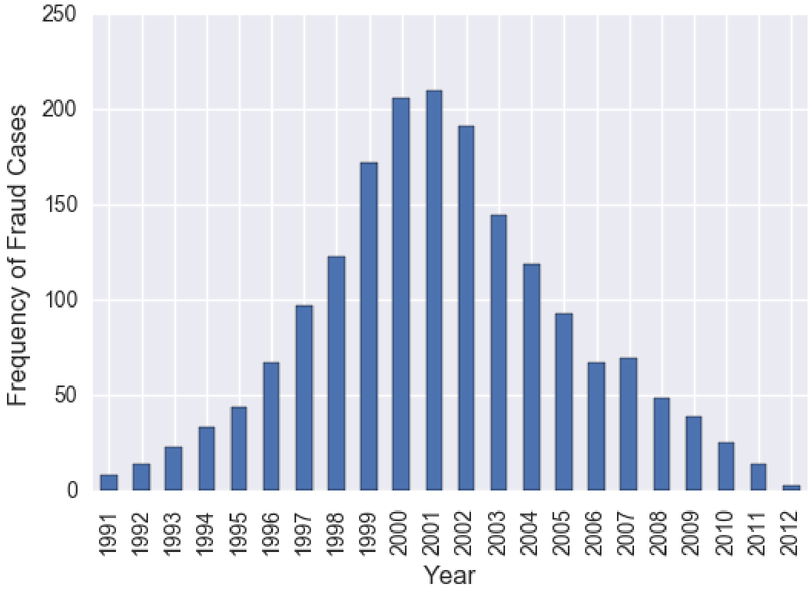
3.6 Exploratory Descriptive Analysis

In order to better understand the data and the phenomenon of interest, an exploratory analysis is conducted including temporal distribution of fraud cases, as well as descriptive statistics visual exploration of the distribution of the collected financial statement information.

It can be observed from Figure 3.1 that most fraud cases were taken place between 1999 and 2002, which coincides with the occurrence of many accounting scandals, such as Enron, Worldcom and Tyco, among others.

In the years following the aforementioned corporate collapses, the number of fraud cases declines significantly. This suggests an important impact of fraud discovery as a discourage mechanism, mainly due to the fear of being discovered as well as the increment of corporate regulation and oversight further enacted to help prevent accounting fraud and malpractices.

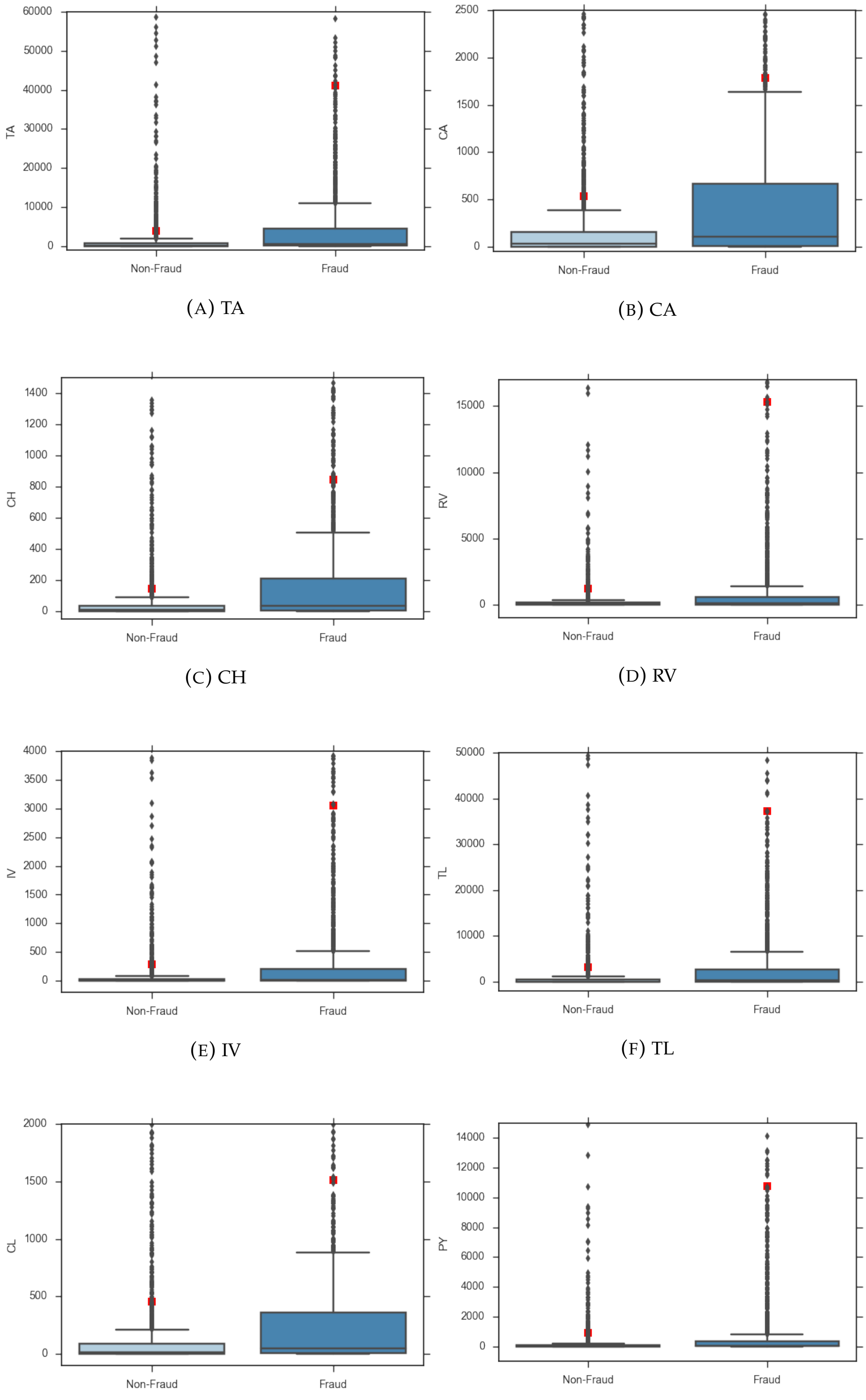
FIGURE 3.1: Fraud cases by year



Interesting characteristics can be noticed from Table 3.2 about the collected dataset. First, large range and standard deviation of total assets stand from the table, which means companies of all sizes are included in the sample. Second, high absolute values of skewness and kurtosis suggest asymmetric distribution of most financial items, which is quite expected as companies of all sizes and industries have been considered.

Furthermore, boxplots of all financial items have been constructed in order to explore the relationship between them and the response variable *Fraud*.

Presence of outliers can be seen in most cases, for both fraud and non-fraud groups. In addition, higher variation in the case of fraudulent firms is observable in several financial items, including CA, CL, TE, NI, SA and COGS. The distribution of the data regarding TA, RV, IV, TL, PY, RE and EBIT appears to be similar in both groups. More about distribution of the data and important difference between fraud and non-fraud cases will be extensively discussed in the next chapter.



(A) TA

(B) CA

(C) CH

(D) RV

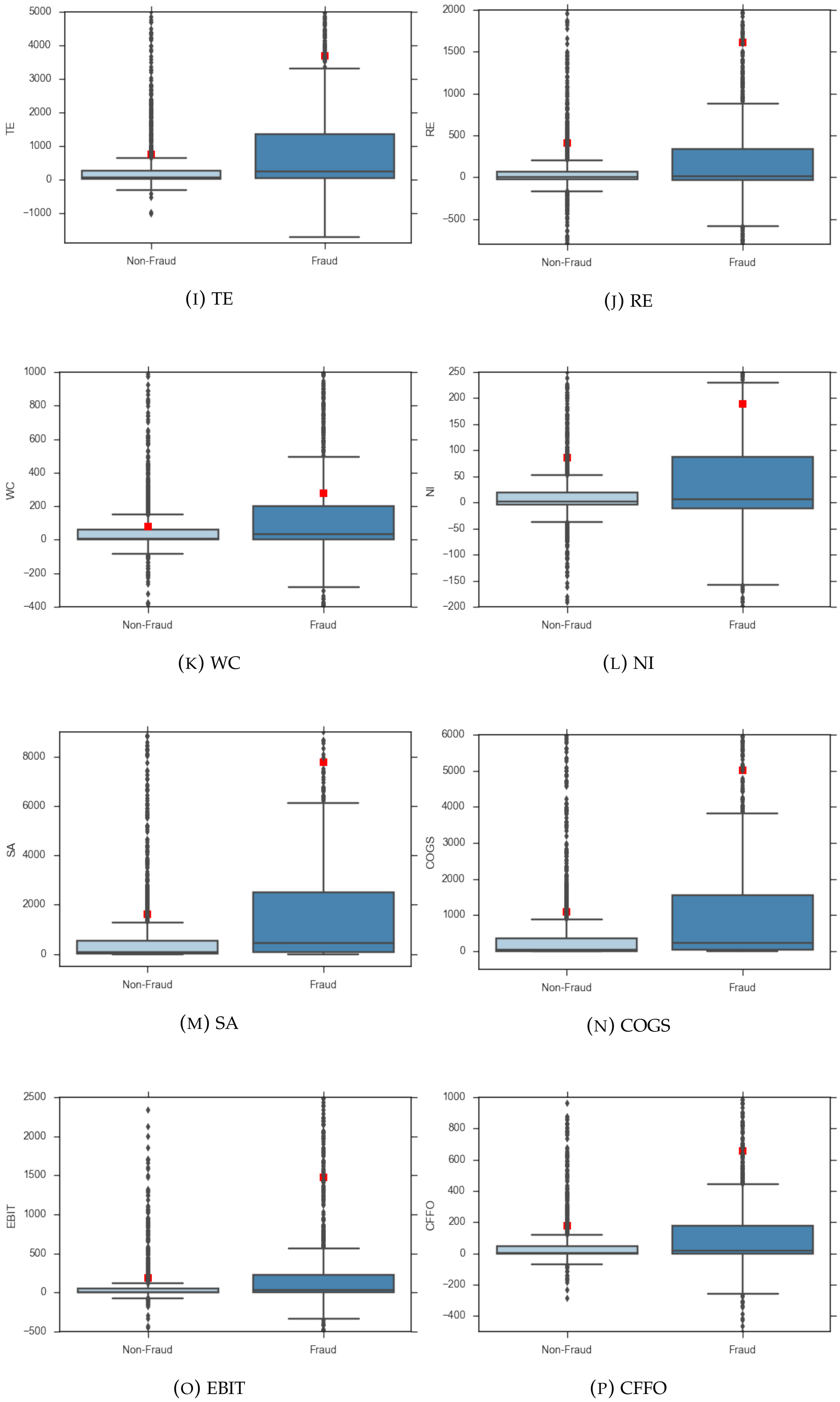
(E) IV

(F) TL

(G) CL

(H) PY

FIGURE 3.2: Boxplots of examined financial items for non-fraud and fraud firms



(I) TE

(J) RE

(K) WC

(L) NI

(M) SA

(N) COGS

(O) EBIT

(P) CFFO

FIGURE 3.2: Boxplots of examined financial items for non-fraud and fraud firms (continued)

TABLE 3.2: Descriptive statistics of selected financial statement items

Item	Mean	Std. Dev.	Min	Max	Skewness	Kurtosis
TA	22495.31	153227.40	0	2359141	10.565	128.601
CA	1160.71	4845.78	0	96853	9.161	114.911
CH	496.01	2890.65	0	59602	11.958	177.477
RV	8260.99	63665.26	0	994847	10.580	125.770
IV	1669.36	20832.87	0	472266	18.487	364.381
TL	20177.11	144472.90	0	2155072	10.516	126.474
CL	983.73	4751.81	0	111604	10.962	171.183
PY	5822.05	59749.21	0	1193593	14.153	220.468
LTD	4421.26	32539.25	0	486876	10.955	135.037
TE	2218.39	10902.03	-30731	204069	11.163	160.561
RE	1016.95	6782.13	-55548	117260	8.345	113.133
WC	176.98	2367.02	-111604	19877	-32.178	1568.497
NI	137.94	2191.30	-58707	21284	-11.170	308.495
SA	4687.70	18001.12	0	297107	7.471	76.364
COGS	3052.69	12788.18	0	245165	8.916	110.310
EBIT	834.39	4881.48	-10537	88847	9.361	108.142
CFFO	421.12	4981.35	-110560	121897	3.335	286.323

3.7 Summary

In Chapter 3, a comprehensive analytical approach is proposed to properly gather and prepare a representative sample of fraud and non-fraud cases, as well as relevant financial information related to both fraudulent and non-fraudulent firms. After the collection process, several tasks associated with the manipulation of raw data are taken place, including data cleaning, data transformation, data merging, data validation and missing values treatment.

Finally, a clean, organised and structured database is achieved and analysed, which leads to the next stage of defining and selecting potential explanatory variables to be used to accurately detect accounting fraud offences. In particular, financial ratios will be proposed as explanatory variables of accounting fraud and further assessed in terms of detection power.

Chapter 4

Financial Ratio Analysis

It is inspiring to witness all the great work researchers have done by virtue of accounting fraud detection, not just in terms of statistical analysis and modelling, but also in regards to the input variables used as predictors of fraudulent reporting. Nevertheless, there is no consensus on which data features are best for detecting corporate wrongdoing, probably due to the subjective nature of financial reporting and the always-evolving dynamic of this type of crime.

Although the use of analytical procedures has certainly increased the effectiveness of accounting fraud detection (Section 2.7), the analysis of more meaningful information would undeniably help to achieve more accurate results. It is reasonable to think that changes in aggregate cycles or the relationship between different financial items will have a greater explanatory power as opposed to individual account information (Choi and Green, 1997). Hence, it is generally believed that financial ratio data is more effective than accounting data, especially when noticing that it leads to noticeably better results and predictions (Wang, 2010).

A ratio expresses two values or measurements relative to each other, and it is a very convenient figure since it facilitates the comparison between the two quantities of interest. One of the advantages of using this kind of calculation is the straightforward interpretation of a ratio, that is indeed, the number of times that the numerator contains or is contained within the denominator.

That being said, and taking into account the overwhelming amount of information contained in financial reports, then a smart selection of relevant financial ratios is required, considering accounting items that may be more susceptible of being manipulated and that properly identify key aspects of a firm.

First, financial ratios will be defined and justified in accordance of literature popularity, as well as the expected relationship between them and the target variable *Fraud*. Then, an exhaustive analysis of financial ratios will be performed employing the entire collected database, regardless of the economic sector the company belongs to. At last, a more thorough examination will be conducted by industry to further explore whether there are different domain-specific accounting tricks executives tend to use to commit fraud.

In order to construct useful financial ratios and properly analyse them, financial statement information of public U.S. companies will be used (Chapter 2). The transformation of the financial data into valuable ratio information is conducted in the following section, as well as an in-depth analysis of the most relevant ones in terms of explanatory capability.

4.1 Financial Ratios

As mentioned before, a great deal of research studies includes subjective judgment and/or qualitative and non-public information into their models, that are only available to auditors and insiders of the sampled firms. Accounting data, on the other hand, is publicly available for external interested parties, hence whether it can be used to detect falsified reporting is an intriguing question (Persons, 1995).

The literature suggests that financial statement information is useful for accounting fraud detection. In particular, it can be seen that ratio analysis is very popular for this end suggesting that a careful reading of financial ratios can reasonably expose symptoms of fraudulent behaviour. As such, ratios are calculated to quantify the relation between two financial items and to subsequently define acceptable non-fraudulent values. Therefore, if a fraudulent activity is taking place, financial ratios associated with manipulated accounts will deviate from the normal behaviour and conveniently exhibit signs of accounting fraud.

Although the usefulness of financial ratios has been recognised by many researchers along the years, they also exhibit problems related to near-zero denominators and dissimilar signs of numerators and denominators. When facing this kind of issues, the obvious and most tempting option is to delete the particular observation or company from the analysis, which has been previously done in Section 3.4.4 that deals with the data validation process.

There has been an interesting debate (Section 2.7) about which features should be used for detecting falsified reports, but still no agreement on which ones are best for this end. An in-depth analysis of the most severe accounting scandals occurred in the U.S. in the last few decades (Shilit and Perler, 2010) shows that the most frequent tricks managers employ in order to hide debilitated businesses are commonly associated with the manipulation of earnings and cash flow items.

In this manner, and considering relevant and significant variables resulting from prior research work on the topic, this study identifies 20 financial statement ratios that measure the majority of aspects of a firm's financial performance, including leverage, profitability, liquidity and efficiency. All financial ratios are computed using the selected financial items previously described in Section 3.2.

Leverage

One of the most important aspects of a firm is leverage, since it represents the potential return of an investment based on the debt structure of the company. When debt is used to purchase assets then the value of assets exceeds the borrowing cost, basically because debt interest is tax deductible. However, this practice comes with greater risks for investors, considering that sometimes firms are not able to pay their debt obligations.

In consequence, companies having trouble paying their debts may be tempted to manipulate financial statements in order to meet debt covenants. Therefore, high levels of debt should increase the likelihood of accounting fraud, since it transfers the risk from the firm and its managers to shareholders.

Many studies have measured this aspect using the following financial ratios:

- Total Liabilities to Total Assets (TLTA): It has been shown that debt size compare to the total value of a company is a significant metric to assess accounting fraud, since the higher this proportion is, then less risk is taken by the equity owners and managers and more risk is shifted to investors (Persons, 1995; Spathis, 2002; Spathis, Doumpos, and Zopounidis, 2002; Kaminski, Wetzel, and Guan, 2004; Kotsiantis et al., 2006; Lenard, Watkins, and Alam, 2007; Pai, Hsu, and Wang, 2011; Song et al., 2014). Therefore, it is expected to find higher levels of leverage in fraudulent firms than non-fraudulent and, in consequence, a positive relation between TLTA and fraud.

- Total Liabilities to Total Equity (TLTE): As before, a positive association is expected between TLTE and fraud, as managers may be tempted to increase debt in order to reduce the risk of equity owners when facing difficult times. The inclusion of this variable seems reasonable considering the evidence of its power to detect accounting fraud (Fanning and Cogger, 1998; Kirkos, Spathis, and Manolopoulos, 2007; Dalnial et al., 2014).
- Long-Term Debt to Total Assets (LTDTA): It has been suggested that since the estimation of accounts related to long-term obligations is subjective, then it is easier to manipulate them (Kirkos, Spathis, and Manolopoulos, 2007; Pai, Hsu, and Wang, 2011). Consequently, the more difficult is to detect falsified long-term items and the more attractive is to commit fraud using these accounts. Again, it is expected to find higher levels of LTDTA in fraudulent firms compared to non-fraudulent firms.

Profitability

Profitability measures are used to estimate a firm's ability to generate earnings compared to its costs, hence the importance of maintaining these metrics in line with market projections. As consequence, executives may be willing to manipulate earnings-related financial statements in order to cover profitability problems when companies are not performing as expected.

To test whether firms with poorer financial condition are more likely to engage in fraudulent financial reporting, relevant ratios associated with income, expenses and retained earnings will be considered and accordingly, define below.

- Net Income to Total Assets (NITA): Many studies have exposed the utility of using NITA as an explanatory variable of accounting fraud (Persons, 1995; Spathis, 2002; Spathis, Doumpos, and Zopounidis, 2002; Kaminski, Wetzel, and Guan, 2004; Kirkos, Spathis, and Manolopoulos, 2007; Lenard, Watkins, and Alam, 2007; Ravisankar et al., 2011; Pai, Hsu, and Wang, 2011; Gupta and Gill, 2012; Dalnial et al., 2014; Song et al., 2014). The rationale behind the use of this ratio is that when profit projections are not met, then overstating revenue or understating expenses may be a practical solution. In consequence, it would not be surprising to see unusually high levels of income compare to the size of the business, i.e.: total assets, when it comes to fraudulent companies.

- Retained Earnings to Total Assets (RETA): Retained earnings make direct reference to accumulated profit, so similarly to the previous ratio, a positive relation is expected between RETA and fraud occurrence considering management temptation of maliciously exaggerate these records in order to please shareholders. Prior research work support this theory as it can be seen in Lee, Ingram, and Howard (1999), Kaminski, Wetzel, and Guan (2004), and Gupta and Gill (2012).
- Earnings Before Interest and Tax to Total Assets (EBITTA): EBIT, also referred as *operating income*, is one of the most important indicators of a company's profitability and, consequently, very likely to be modified to further hide mediocre performance. It is believed that fraudulent firms are prone to improperly magnify this particular financial item when needed, hence the usefulness of comparing this metric to total assets as predictor of accounting fraud (Kotsiantis et al., 2006).

Liquidity

Liquidity refers to the ability to which an asset can be converted from an investment to cash. This concept is highly important for businesses and investors, since liquid assets reduce in some extent investing risks by ensuring the capacity of a firm to pay off debts as they come due. Consequently, problems involving liquidity may provide an incentive for managers to commit accounting fraud, hence the need to investigate financial ratios related to the liquid composition of assets, as is the case of working capital and current assets.

- Working Capital to Total Assets (WCTA): By far, one of most popular liquidity ratios used to predict accounting fraud probably because of its importance for shareholders (Spathis, 2002; Spathis, Doumpos, and Zopounidis, 2002; Kaminski, Wetzel, and Guan, 2004; Kotsiantis et al., 2006; Kirkos, Spathis, and Manolopoulos, 2007; Pai, Hsu, and Wang, 2011). As define in Section 3.1, working capital is the difference between current assets and current liabilities, representing the capital needed by a company to successfully perform its daily operations. Lower liquidity will encourage executives to inflate current assets and eventually overstate working capital, then fraudulent firms should show higher values for WCTA.

- Current Assets to Total Assets (CATA): Some studies have used current assets directly (Persons, 1995; Lenard, Watkins, and Alam, 2007) based on the claim that the higher this value as a proportion of total assets, the more likely is that a company is committing accounting fraud. Hence, a positive association with this ratio is expected.
- Current Assets to Current Liabilities (CACL): Also known as *current ratio*, this metric is mainly used to determine company's financial health, since it measures the ability to pay short-term debt and other payables, i.e.: current liabilities. Acceptable values of CACL vary depending on the industry, but it is believed that dishonest firms will tend to exaggerate this ratio as much as possible to ultimately project a favourable economic position (Kotsiantis et al., 2006; Lenard, Watkins, and Alam, 2007; Ravisankar et al., 2011; Song et al., 2014).
- Cash to Net Income (CHNI): A very important component of current assets is cash as it represents the company's most immediate instrument of exchange. Alteration of this item is somewhat difficult due to its tangible nature, hence increase in net income and not in cash may be an indicator of accounting fraud. Therefore, lower values of CHNI are expected in presence of irregular activities.

Many investors have alternatively focused their attention on the company's capability to generate cash from its actual business operations. This aspect however, is usually manipulated since "companies can exert a great deal of discretion when presenting cash flows" (Shilit and Perler, 2010). Ergo, the importance of thoroughly analyse cash flow from operations and, in particular, evaluate its relationship with reported earnings.

- Cash Flow From Operations to Net Income (CFFONI): Cash flow from operations and net income are valid metrics of businesses' performance, hence the expectation of both moving in the same direction, that is, systematic change in one of them shall be accompanied by a similar change in the other. Therefore, disparities between these items should be taken seriously, as it may indicate that accounting fraud is being perpetrated. Fraudulent firms may try to increase earnings but fail to boost CFFO levels, so lower values of this ratio are expected in these instances.

Efficiency

Financial efficiency refers to the capacity of producing as much as possible using as few resources as possible. Inefficiency usually involves higher costs, hence resulting in poorer firm's performance, which may motivate managers to misstate financial statements that allow subjective estimations, and therefore, are easier to manipulate. Such is the case of accounts receivable, accounts payable, inventory and cost of good sold, so financial ratios related to these accounts are further selected.

- Accounts Receivable to Total Sales (RVSA): There is strong evidence of the significance of this ratio when detecting accounting fraud (Fanning and Cogger, 1998; Feroz et al., 2000; Kaminski, Wetzel, and Guan, 2004; Pai, Hsu, and Wang, 2011). Many cases of dishonest reporting involve the inflation of current-period earnings through the incorrect early recognition of revenue or the recognition of fictitious earnings (Shilit and Perler, 2010). A clear sign of these strategies is when accounts receivable grow much faster than sales, hence the need to assess the relationship between these two variables and be careful when this ratio is smaller than usual.

- Accounts Receivable to Total Assets (RVTA): Accounts receivable are also a component of current assets (CA), along with cash and inventories. Thereby, in order to artificially increase CA, managers may be tempted to exaggerate receivables, in particular considering how difficulty is to audit this kind of transactions. Then, a positive relation is expected as fraudulent firms should have higher levels of accounts receivable compare to total assets than non-fraudulent firms (Lin, Hwang, and Becker, 2003; Kaminski, Wetzel, and Guan, 2004; Kotsiantis et al., 2006).

- Inventory to Total Sales (IVSA): It is believed that weakened companies usually have lower stock turnover with respect to sales (Fanning and Cogger, 1998; Spathis, 2002; Spathis, Doumpos, and Zopounidis, 2002; Pai, Hsu, and Wang, 2011). Large inventories may make the firm more vulnerable to accounting fraud, so a positive relation is expected between IVSA and fraud occurrence.

- Inventory to Total Assets (IVTA): Research findings suggest that fraudulent companies tend to maliciously overstate inventories to boost current assets (CA) and ultimately hide business deterioration (Ravisankar et al., 2011; Gupta and Gill, 2012; Dalnial et al., 2014; Song et al., 2014). Legitimate increase in inventories should be naturally followed by an increment in total assets, hence high values of IVTA should be taken seriously, as it may be an indicator of dishonest valuation of inventories.

- Inventory to Current Assets (IVCA): Again, higher levels of inventories directly compare to current assets may be suggesting improper financial estimates of inventory (Ravisankar et al., 2011)
- Inventory to Cost of Good Sold (IVCOGS): Typically, converting inventory into expense happens immediately after a sale has occurred. Nevertheless, in some cases it is not such a straightforward process, and unethical managers may be temped to take advantage of this situation and intentionally fail to record necessary expenses for excess and obsolete inventory. This could potentially lead to artificial high values of inventory in addition to low values of cost of good sold. For this reason, it is expected that fraudulent firms show higher values of IVCOGS compared to non-fraudulent businesses (Kaminski, Wetzel, and Guan, 2004; Ravisankar et al., 2011; Song et al., 2014).
- Accounts Payable to Cost of Good Sold (PYCOGS): A common technique used to commit accounting fraud is the increment of operating cash flow through unsustainable activities, in particular, boosting CFFO by artificially decreasing outstanding accounts (Shilit and Perler, 2010). In particular, the ratio of payables to cost of good sold describes how much the company has as pending payments in terms of operating expenses, so lower values of PYCOGS may be the result of the aforementioned technique, and in consequence, a clear sign of accounting fraud.

Efficiency it also linked to capital turnover, which represents the sales generating power of a firm's assets. In order to maintain the appearance of consistent growth, fraudulent managers may be tempted to manipulate sale-related financial items when dealing with competitive situations. Accordingly, the following two sale-ratios are considered in order to identify possible fictitious trend in growth.

- Total Sales to Total Assets (SATA): Unjustified jumps in revenue that are not in harmony with the size of the company (i.e.: total assets) should be always taken in consideration when examining financial reports since it is often related to accounting fraud (Fanning and Cogger, 1998; Spathis, Doumpos, and Zopounidis, 2002; Kotsiantis et al., 2006; Kirkos, Spathis, and Manolopoulos, 2007; Lenard, Watkins, and Alam, 2007). Therefore, a positive association between SATA and fraud occurrence is awaited.

- Total Sales to Total Equity (SATE): Similarly, increments in revenue but not in shareholders' equity is rather suspicious considering that both items should move together in the same direction. Hence the need to evaluate this ratio to look for accounting fraud symptoms.

A detailed analysis of the aforementioned ratios will be performed next to evaluate the best explanatory variables of accounting fraud and the relationship between each other and also in regards to the target variable, *Fraud*. But first, detection and treatment of extreme values has to be done to remove uninformative observations that may distort the analysis and further modelling of the studied fraudulent behaviour.

TABLE 4.1: Summary of considered financial ratios and calculation

Category	Financial Ratio	Calculation
Leverage	TLTA	Total Liabilities / Total Assets
	TLTE	Total Liabilities / Total Equity
	LTDTA	Long-Term Debt / Total Assets
Profitability	NITA	Net Income / Total Assets
	RETA	Retained Earnings / Total Assets
	EBITTA	Earning Before Interest and Tax / Total Assets
Liquidity	WCTA	Working Capital / Total Assets
	CATA	Current Assets / Total Assets
	CACL	Current Assets / Current Liabilities
	CHNI	Cash / Net Income
	CFFONI	Cash Flow From Operations / Net Income
Efficiency	RVSA	Accounts Receivable / Total Sales
	RVTA	Accounts Receivable / Total Assets
	IVSA	Inventory / Total Sales
	IVTA	Inventory / Total Assets
	IVCA	Inventory / Current Assets
	IVCOGS	Inventory / Cost of Good Sold
	PYCOGS	Accounts Payable / Cost of Good Sold
	SATA	Total Sales / Total Assets
	SATE	Total Sales / Total Equity

4.2 Outlier Detection

Outliers or extreme observations are values that are very distant from other observations, or in simple words, that are unusually small or large compared to the rest of the dataset. There are two types of outliers, the ones that occur by chance and are considered to be valid, and the ones that result from measurement or recording errors typically considered as invalid observations. Regardless of whether they are valid or not, it is imperative to understand how they may affect the analysis, in particular when financial ratios are being evaluated.

As mentioned above, financial ratios are very useful when comparing two numerical values, but suffer from a significant disadvantage. Given that ratios are fractions, then zero denominators or values close to zero may result in undefined expressions or extremely large values, leading to meaningless comparisons between the studied financial items. In these cases, outliers may be very influential and potentially cause important distortions in terms of descriptive statistics, inferential analysis and statistical modelling. Hence the importance of carefully locate them and later extract them using an appropriate analytical approach.

In general, there are two ways of dealing with outliers. The first approach uses basic descriptive statistics and useful visual mechanisms to detect extreme observations, such as minimum and maximum values, means and variances, as well as histograms, boxplots and scatterplots.

The second approach instead, locates outliers calculating how far the observations are from the centre, that is, how many standard deviations each data value lies away from the mean. For this end, z-scores are calculated for all the observations using the following formula:

$$Z = \frac{X - \bar{X}}{S} \tag{4.1}$$

where X represents the data value, $\bar{X}$ is the sample mean and S the sample standard deviation.

It is commonly accepted that if an observation is situated three standards deviations away from the mean, then is considered as an outlier. Or in other words, if the absolute value of its z-score is bigger than three. This method is finally adopted in order to detect and remove extreme values present in the database.

4.3 Ratio Analysis

In order to perform a comprehensive analysis of the selected financial ratios previously defined, a summary of them is presented in Table 4.8, along with the involved mathematical calculation and the expected relationship with regard to accounting fraud.

TABLE 4.2: Summary of selected financial ratios and expected relation with target variable *Fraud*

Financial Ratios	**Expected Relationship**
TLTA	Positive
TLTE	Positive
LTDTA	Positive
NITA	Positive
RETA	Positive
EBITTA	Positive
WCTA	Positive
CATA	Positive
CACL	Positive
CHNI	Negative
CFFONI	Negative
RVSA	Negative
RVTA	Positive
IVSA	Positive
IVTA	Positive
IVCA	Positive
IVCOGS	Positive
PYCOGS	Negative
SATA	Positive
SATE	Positive

To evaluate significant differences between financial accounts related to fraudulent and genuine reports, two hypothesis testing techniques will be described and implemented next. First, a parametric approach will be performed to test differences in the average values of financial ratios of fraud and non-fraud firms. Then, a non-parametric method will be implemented to test if the distribution of fraudulent data differs significantly compared to non-fraudulent data.

Two Sample t-test

A simple yet very informative univariate analysis is performed as a first step to understand key financial indicators that may be suggesting that accounting fraud has been or is being committed. In particular, t-tests are conducted to assess whether the examined financial ratio has the same mean within the two groups, that is, non-fraud and fraud firms. Therefore, the following hypotheses are specified:

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 \\ H_1 &: \mu_1 \neq \mu_2 \end{aligned} \tag{4.2}$$

where the subscript 1 makes reference to non-fraud firms and the subscript 2 makes reference to fraud firms.

It can be observed (Table 4.3) that sample standard deviations of both groups are reasonably dissimilar for all selected financial ratios. Therefore, it will be assumed that both populations, non-fraud and fraud, have unequal variances and, consequently, the following formula for the t-statistic corresponding to the hypothesis testing specified above will be considered:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)}} \tag{4.3}$$

where $\bar{X}$ represents the sample mean, s^2 is the sample variance and n is the sample size. Again, subscript 1 makes reference to non-fraudulent firms and subscript 2 to fraudulent companies.

The resulting statistic is distributed as Student's t with ν degrees of freedom, where ν is given by:

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\dfrac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \dfrac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} \tag{4.4}$$

After the test has been performed, the *p-value* associated with the corresponding statistic can be easily calculated as the probability of obtaining a result equal to or more extreme than the one obtained from the sample considering that the null hypothesis is true. If this probability is low, then it is very unlikely that a value as extreme as the one obtained from the sample is observed, hence it seems suspicious that the null hypothesis is true. Accordingly, if *p-value* is lower than the significance level[1], then the null hypothesis H_0 can be rejected so the evidence favours the alternative, H_1. Therefore, it can be said that there is a significant difference between the non-fraudulent firms and fraudulent firms.

Moreover, when saying that there is a significant difference between the groups, it means that the analysed financial ratio has the power to explain in some degree the variable of interest, that is, the target variable *Fraud*. Hence it can be used as meaningful information for accounting fraud detection. If the financial ratio is not significant, it will not contribute much to the analysis and, in consequence, it makes no sense to include it as a predictive variable.

It is worth mentioning the importance of interpretability when adopting statistical techniques for detecting accounting fraud. In addition to the tests that will be performed next, the expectation of the relation between the studied ratios and the target variable will also be analysed, as it is a very straightforward task just by inspecting which group, non-fraud or fraud, shows a larger mean. It is hereby highly preferable to meet what was anticipated by business experts and researchers, otherwise potential users will be reluctant to use the proposed methodology.

Results of the proposed 20 tests are shown in Table 4.3. All relevant information of the testing approach is presented, including sample mean, sample standard deviation and sample size of both groups, as well as the resulting t-statistics, degrees of freedom and p-values for all selected financial ratios.

An exhaustive analysis of these results will be performed next, examining whether the included financial ratios are statistically significant and in accordance with the expectations described before.

[1] A significance level of 0.05 will be considered for this test.

TABLE 4.3: T-tests for the difference in the mean of non-fraud and fraud firms for the 20 financial ratios

Ratio	Mean*		Standard Deviation*		Sample Size		t-stat**	Degrees of Freedom	p-value (two-tailed)
	Non-Fraud	*Fraud*	*Non-Fraud*	*Fraud*	*Non-Fraud*	*Fraud*			
TLTA	0.5995	0.5984	0.4472	0.3873	215,860	1,594	0.1172	1,624.5	0.9067
TLTE	2.2421	2.7503	7.4124	7.7229	215,860	1,594	-2.6184	1,614.8	0.0089
LTDTA	0.1821	0.1864	0.2514	0.2206	215,425	1,588	-0.7767	1,617.5	0.4375
NITA	0.2002	-0.1283	23.4581	0.6068	215,860	1,594	-1.3648	121,544.0	0.1723
RETA	-2.2266	0.7108	53.6623	4.3381	210,312	1,577	-9.4689	7,196.3	0.0000
EBITTA	-0.1182	-0.0316	4.1750	0.4467	210,869	1,582	-5.9906	4,317.6	0.0000
WCTA	0.1451	0.1660	0.3953	0.3587	215,860	1,594	-2.3173	1,621.7	0.0206
CATA	0.4067	0.4376	0.3116	0.2894	215,860	1,594	-4.2537	1,620.4	0.0000
CACL	3.9860	2.5758	63.2358	2.8209	180,130	1,340	8.4075	27,233.1	0.0000
CHNI	1.8151	0.2032	219.2324	28.7837	210,600	1,566	1.8523	3,202.0	0.0641
CFFONI	1.5170	0.0596	62.9141	57.5945	200,395	1,556	0.9936	1,584.0	0.3206
RVSA	1.5153	0.8701	50.2736	2.3945	204,968	1,582	5.1082	28,129.7	0.0000
RVTA	0.1967	0.2007	0.2047	0.1723	215,860	1,594	-0.9174	1,626.4	0.3591
IVSA	0.2719	0.2684	14.7204	1.7448	204,968	1,582	0.0646	3,786.4	0.9485
IVTA	0.0974	0.1178	0.1435	0.1565	215,860	1,594	-5.1946	1,612.8	0.0000
IVCA	0.2083	0.2285	0.2212	0.2286	178,379	1,339	-3.2283	1,356.9	0.0013
IVCOGS	0.3528	0.5381	9.1617	4.2733	206,952	1,582	-1.6958	1,694.0	0.0901
PYCOGS	3.4372	1.7469	27.1332	6.5998	205,061	1,565	9.5354	1,993.3	0.0000
SATA	0.9248	0.9655	1.9753	0.8745	215,860	1,594	-1.8271	1,715.3	0.0679
SATE	2.2589	2.5098	11.7867	6.9577	215,860	1,594	-1.4243	1,661.2	0.1545

Notes:

* The amounts are reported in million USD

** Two-sample t test with unequal variances

1. **TLTA**: There is virtually no difference between non-fraud and fraud firms, hence no apparent contribution from this financial ratio.
2. **TLTE**: The *p-value* of this ratio suggests a significant effect on the target variable *Fraud*, and sample means are in agreement with the expectation, that is, fraudulent firms show higher levels of liabilities compared to shareholders' equity than non-fraudulent companies.
3. **LTDTA**: Slightly higher mean for the fraudulent group, as expected, but not statistically significant.
4. **NITA**: Lower, and negative, mean of net income to total assets for fraud firms, which is contradictory to the expected. Not significant explanatory variable, whatsoever.
5. **RETA**: Important discrepancy between the average of both groups (Figure 4.1a), higher for fraudulent companies as expected.
6. **EBITTA**: Significant higher levels of EBIT compared to total assets can be observed for fraudulent companies (Figure 4.1b), suggesting managers' preference for manipulating earnings figures.
7. **WCTA**: The higher the value of working capital compared to total assets, the higher the likelihood of committing accounting fraud.
8. **CATA**: A highly significant positive influence of this financial ratio is suggested by the *p-value*, which is in harmony with prior expectation.
9. **CACL**: It can be seen from Figure 4.1c that fraudulent firms present significantly lower levels of CACL compared to non-fraudulent companies, which is contradictory to what was expected.
10. **CHNI**: The expectation of lower values of cash compared to net income for dishonest firms is supported by Figure 4.1d, although not in a significant way.
11. **CFFONI**: Lower levels of CFFO to net income can be observed for the fraudulent group as expected, but not significant enough.
12. **RVSA**: As it can be seen in Figure 4.1e, fraudulent firms have, on average, lower values of RVSA compared to non-fraudulent companies. This result is statistically significant and in accordance to what was expected.

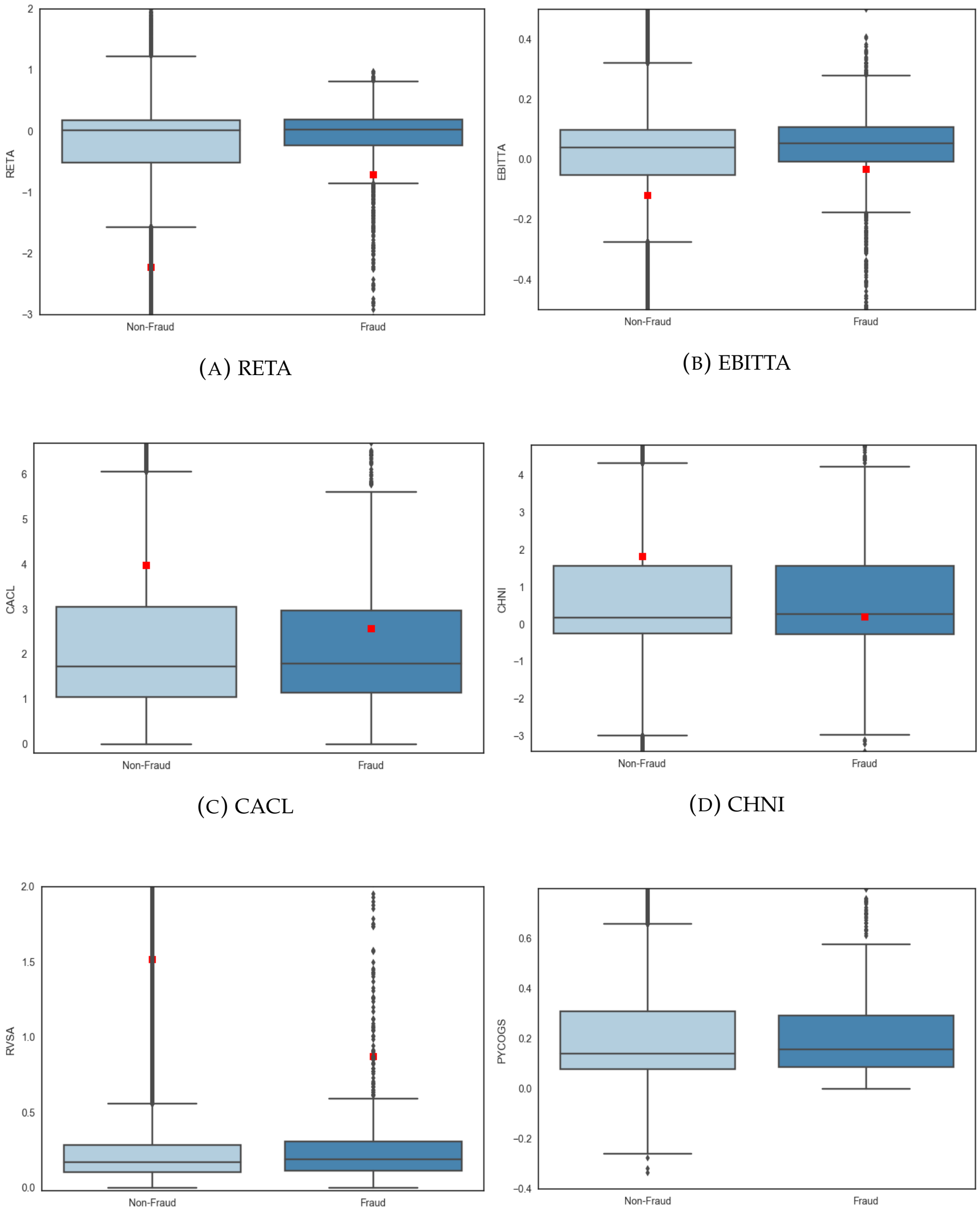
13. **RVTA**: There is no clear difference between non-fraud and fraud firms in terms of RVTA, thus no significant contribution made by this ratio.
14. **IVSA**: Virtually no difference between groups, hence not significant predictor.
15. **IVTA**: Significantly higher values of IVTA within fraudulent firms, supporting the suspicion of artificially exaggerate inventory levels as a means to commit accounting fraud.
16. **IVCA**: Same as before, and as expected, significantly higher levels of inventory to current assets is found in fraud cases.
17. **IVCOGS**: The higher this financial ratio, the more likely it is that a company is committing fraud. Significant result when one-tailed test is considered.
18. **PYCOGS**: Significantly lower levels of accounts payable compare to cost of good sold were found for fraudulent firms, as expected (Figure 4.1f).
19. **SATA**: In accordance with the expectation, a significant positive effect of this financial ratio is suggested when one-side test is considered.
20. **SATE**: Slightly higher mean for the fraudulent group, as expected, but not statistically significant.

Mann-Whitney test

An alternative hypothesis testing technique is proposed as t-test may be suffering from important drawbacks. The so-called Mann-Whitney test is a non-parametric method that is commonly employed due to its ease of use and availability in several advanced statistical software.

In simple terms, non-parametric methods refer to statistical techniques that do not make assumptions on the data distribution, hence the reason they are also called distribution-free tests (Hollander, Wolfe, and Chicken, 2013). These models are particularly useful when there are definite outliers or extreme observations in the data, as is the case of the studied database.

The Mann-Whitney test is performed using the rank of the data, that is, the position of each observation within the sample rather than the value *per se*. In light of this, then it is easy to notice that outliers will have a minimal effect on the test, which makes it very robust in terms of extreme values (Sheskin, 2003).



(A) RETA

(B) EBITTA

(C) CACL

(D) CHNI

(E) RVSA

(F) PYCOGS

FIGURE 4.1: Boxplots of Significant Financial Ratios for Non-fraud and Fraud Firms

The test procedure starts calculating the rank of every data point in each sample, fraud and non-fraud firms. Ranks are ordered first within the first group (fraud firms) and then within the second (non-fraud firms), and compared later using a test statistic that measures the number of ranking discrepancy between both groups. If groups are similar, ranks will look alike and, in consequence, the distribution of the data for both samples should be equivalent. On the contrary, if groups are dissimilar, then ranks will differ, which suggests different distribution across the groups.

The following hypotheses are specified for the Mann-Whitney test:

$$\begin{aligned} H_0 &: \textit{the distribution of both groups are equal} \\ H_1 &: \textit{the distribution of both groups are not equal} \end{aligned} \tag{4.5}$$

Same as t-tests, the test statistic and *p-value* are calculated for all financial ratios. Accordingly, if *p-value* is lower than the significance level[2], then the null hypothesis H_0 can be rejected so the evidence favours the alternative, H_1. Therefore, it can be said that there is a significant difference between the non-fraudulent firms and fraudulent firms with regard to the financial ratio of interest.

Results of the 20 Mann-Whitney tests are shown in the table below:

TABLE 4.4: Two-sample Mann-Whitney test

Ratios	statistic*	p-value
TLTA	-1.979	0.0479
TLTE	-3.952	0.0001
LTDTA	-4.789	0.0000
NITA	-1.382	0.1669
RETA	-4.067	0.0000
EBITTA	-5.858	0.0000
WCTA	-3.575	0.0004
CATA	-3.961	0.0001
CACL	-1.979	0.0479
CHNI	-1.185	0.2360
CFFONI	4.280	0.0000
RVSA	-4.242	0.0000
RVTA	-6.640	0.0000
IVSA	-6.405	0.0000
IVTA	-7.740	0.0000
IVCA	-3.907	0.0001
IVCOGS	-7.281	0.0000
PYCOGS	-1.858	0.0632
SATA	-6.113	0.0000
SATE	-8.198	0.0000

[2] Again, a significance level of 0.05 will be considered for this test.

It is interesting to see that almost all ratios show *p-values* lower than 0.05, which suggests significant differences between both fraud and non-fraud firms. However, two insignificant ratios are revealed when conducting non-parametric tests, that is, NITA and CHNI. As such, it is decided to exclude them from more advanced statistical modelling due to their poor detection power.

4.4 Ratio Analysis by Industry

The univariate exploration previously conducted, clearly exposed potential associations between the selected financial ratios and the target variable, *Fraud*. As a result of the performed tests, 13 out of the twenty ratios initially considered, were found to be significant, which represents a great first step to further reduce the number of explanatory variables. Nevertheless, assuming that fraudsters behave the same across all sectors is fairly naive, so a more elaborated domain-specific examination is reasonably required. Actually, it soon will be seen that when extending the analysis by industry, interesting patterns emerge from the data.

4.4.1 Standard Industrial Classification Overview

Before performing the proposed industry-specific analysis and modelling, a brief explanation of the different industries is given next, along with a detailed description of subsectors involved in each category.

SIC codes are four-digit numerical representations of major businesses and industries. These codes are assigned based on common characteristics shared in the products, services, production and delivery system of a business or organisation.

1. **Agriculture, Forestry and Fishing**: subdivisions within this industry include agricultural production of crops, livestock and animal specialties, agricultural services, forestry services, fishing, hunting and trapping.

2. **Mining and Construction**: subdivisions within the mining industry include metal and ores mining, mining of nonmetallic minerals, petroleum, drilling oil, and gas exploration and services. The manufacturing industry includes general and heavy construction, building contractors and electrical work, among others.

3. **Manufacturing**: subdivisions within the manufacturing industry include food products and plants, dairy products, fruits, vegetables, food specialities, canned

food, grain and bakery products, sugar products, fats and oils, beverages, tobacco products, mill products, products of fabrics and similar materials, clothing, textile products, wood products and furnitures, papers and allied products, newspapers, books and miscellaneous publishing, chemicals and allied products, plastic materials, pharmaceutical preparations, detergents and cleaning products, cosmetics and sanitation preparations, leather products, glass and stone products, electrical equipment, structural metal products, engine and machinery, electronic, components, computers and devices, motor vehicles and equipment, optical instruments and lenses, surgical and medical instruments and supplies, jewellery, musical instrument, games and toys, and sporting and athletic goods, among others.

4. **Transportation, Communication, Electric, Gas and Sanitary Service**: subdivisions within this industry include local and suburban transit, passenger transportation, trucking and courier services, water and air transportation, transportation services, radio, telephone and telegraph communications, radio and television broadcasting and services, electric, gas and sanitary services transmission and distribution, and water supply, among others.

5. **Wholesale Trade and Retail Trade**: subdivisions within this industry include wholesale goods, supplies, furniture, materials, equipment, hardware and software, retail supply and dealers, department stores, grocery and convenience stores, gasoline stations, clothing and shoe stores, consumer electronic stores, and eating and drinking places, among others.

6. **Financial, Insurance and Real Estate**: subdivisions within the financial industry include national and state commercial banks, saving institutions, deposit banking, credit agencies and institutions, loan and security brokers, and investment advice. The insurance industry includes life insurance, accident and health insurance, hospital and medical service, fire and casualty insurance, and insurance agents, brokers and service. The real estate industry include real estate operators and lessors, operators of buildings, real estate agents, managers and dealers, land developers, real estate investment trusts, and investors, among others.

7. **Services**: subdivisions within this industry include hotels and motels, advertising agencies and services, services to dwellings and other buildings, equipment services, employment agencies, computer programming and data

processing services, business services, automotive repair, service and parking, auto rental and leasing, picture and video production, distribution and rental, gambling transactions, amusement parks, sports and recreational clubs, health services and hospitals, medical laboratories, legal services, educational services, social services, engineering, accounting, research and management services, and consulting services among others.

8. **Public Administration**: this last sector consists of establishments of federal, state and local government agencies that administer, oversee and manage public programs and have executive, legislative or judicial authority over other institutions within a given area.

4.4.2 Analysis by Industry

Twenty t-tests are performed in what follows, one per selected financial ratio, but now considering the sector where sampled companies belong to. Table 4.3 summarises significant predictors and the relationship with the dependent variable for each individual SIC industry. Mann-Whitney tests were also implemented, but it has been decided to omit results mainly due to the similarity with t-test results and to avoid overwhelming the reader with repetitive information.

1. **Agriculture, Forestry and Fishing**: Only four financial ratios were found to be significant in this case. In particular, fraudulent firms belonging to this industry will most probably exhibit higher values of EBITTA, and lower values of CACL, IVSA and PYCOGS when compared to non-fraudulent companies.

2. **Mining and Construction**: Interestingly, fraud companies within this sector are much more aggressive in terms of accounting tricks. It can be seen that 14 out of the initial 20, are very likely to be manipulated. Most of them, including TLTA, TLTE, LTDTA, RETA, EBITTA, RVTA, IVSA, IVTA, IVCA, IVCOGS, SATA and SATE, show higher average values compared to non-fraudulent figures, and lower values when dealing with CACL, RVSA and PYCOGS. It is noticeable how do these companies exaggerate debt obligations, inventory levels and sales in such an obvious way to further distort their financial reports.

3. **Manufacturing**: Numerous gimmicks are also accomplished by deceptive manufacturing companies in order to commit accounting fraud. Some evident indicators of falsified reports in this case include lower values of TLTA, NITA,

RETA, CATA, CACL, CFFONI and RVSA, as well as higher levels of TLTE, EBITTA and IVCA.

4. **Transportation, Communication, Electric, Gas and Sanitary Service**: Most of the manipulated accounts in this industry are associated with the reduction of inventories and sales figures, as well as the exaggeration of earnings. As such, it can be seen lower values of IVSA, IVTA, IVCA, PYCOGS and SATA, and higher values of RETA, again when compared to genuine reports.

5. **Wholesale Trade and Retail Trade**: It seems that the *modus operandi* related to deceptive wholesale and retail traders are connected to artificially increasing retained earnings, current assets and inventory since higher levels of RETA, CATA, IVSA and IVTA are apparent for such deceptive companies.

6. **Financial, Insurance and Real Estate**: In this case, a careful analysis of the selected ratios has to be done considering that several financial items are not available for banks, such as current assets, current liabilities and working capital, among others. That being said, clear signs of accounting fraud can be identified such as higher values of TLTA, TLTE, LTDTA, RETA and IVCOGS, and lower quantities regarding CHNI, RVSA, PYCOGS and SATA.

7. **Services**: Curiously, fraudulent firms in this industry tend to falsified records through the reduction of inventory, accounts payable and sales, and via the exaggeration of earnings. As such, companies committing fraud are likely to present lower values of IVTA, IVA, IVCOGS, PYCOGS, SATA, RVSA and CATA, as well as higher values of RETA and EBITTA.

8. **Public Administration**: In contrast with the previous industry, the Public Administration sector tends to artificially inflate inventory and earnings figures in order to falsify their financial reports. Consequently, fraudulent firms in this industry reveal higher average values of IVSA, IVTA, IVCA, IVCOGS, RETA and EBITTA. It is worth mentioning as well, that higher levels of LTDTA and lower levels of WCTA, CATA and SATA are expected to be shown when accounting fraud is being committed by firms belonging to this domain.

TABLE 4.5: Significant financial ratios by industry

Ratio*	Agriculture, Forestry and Fishing	Mining and Construction	Manufacturing	Transportation, Communications, Electric, Gas and Sanitary Service	Wholesale Trade and Retail Trade	Finance, Insurance and Real Estate	Services	Public Administration
TLTA		+	-			+		
TLTE		+	+			+		
LTDTA		+				+		+
NITA			-					
RETA		+	-	+	+	+	+	+
EBITTA	+	+	+				+	+
WCTA								
CATA			-		+			-
CACL	-	-	-				-	-
CHNI						-		
CFFONI			-					
RVSA		-	-			-	-	
RVTA		+						
IVSA	-			-				+
IVTA		+		-	+		-	+
IVCA		+	+	-	+		-	+
IVCOGS		+				+	-	+
PYCOGS	-	-		-		-	-	
SATA		+		-		-	-	-
SATE		+						

Notes:
+ represents a positive association with the target variable, *Fraud*
- represents a negative association with the target variable, *Fraud*
* Two-tailed test at the 0.05 significance level

There are interesting differences between sectors as some ratios are significant or not depending on the industry the company belongs to. A more detailed analysis and interpretation of these and more results is performed in Chapter 5.

In addition to the ratio analysis previously conducted, an exploration of possible association between the explanatory variables will be performed in order to identify the most relevant financial ratios for detecting accounting fraud.

4.5 Correlation Analysis

A very popular technique, often applied in data analytics, is correlation analysis. This method is used to evaluate possible relationships between numerical variables, which is particularly useful when working with accounting items that inevitably interact with each other due to the composition of a financial statement report.

The correlation coefficient quantifies the direction and strength of the implicit relationship of two variables of interest, and only expresses the association between them, not the causality. Nonetheless, if correlation is found between two variables, then it can be used as an indicator of a potential casual relation.

Correlated features should simultaneously change in accordance to the sign of their relationship. Positive correlations will exist if systematic increase in one variable is followed by an increase in the other. Similarly, negative correlations occur when the increase in one variable leads to the decrease of the other. Additionally, the magnitude of the correlation coefficient indicates the strength of the association. The larger the value, the stronger is the association.

Two different measures of correlation will be used to analyse the relationship between the financial ratios of interest. First, Pearson correlation will be calculated assuming a linear relationship between the ratios, and then Kendall correlations will be computed making no linearity assumption whatsoever.

4.5.1 Pearson Correlation

As a first step, Pearson correlation coefficients will be used to explore possible linear relationship between the explanatory variables. Pearson coefficients always range between $+1$ and -1, and they can be calculated using the following formula:

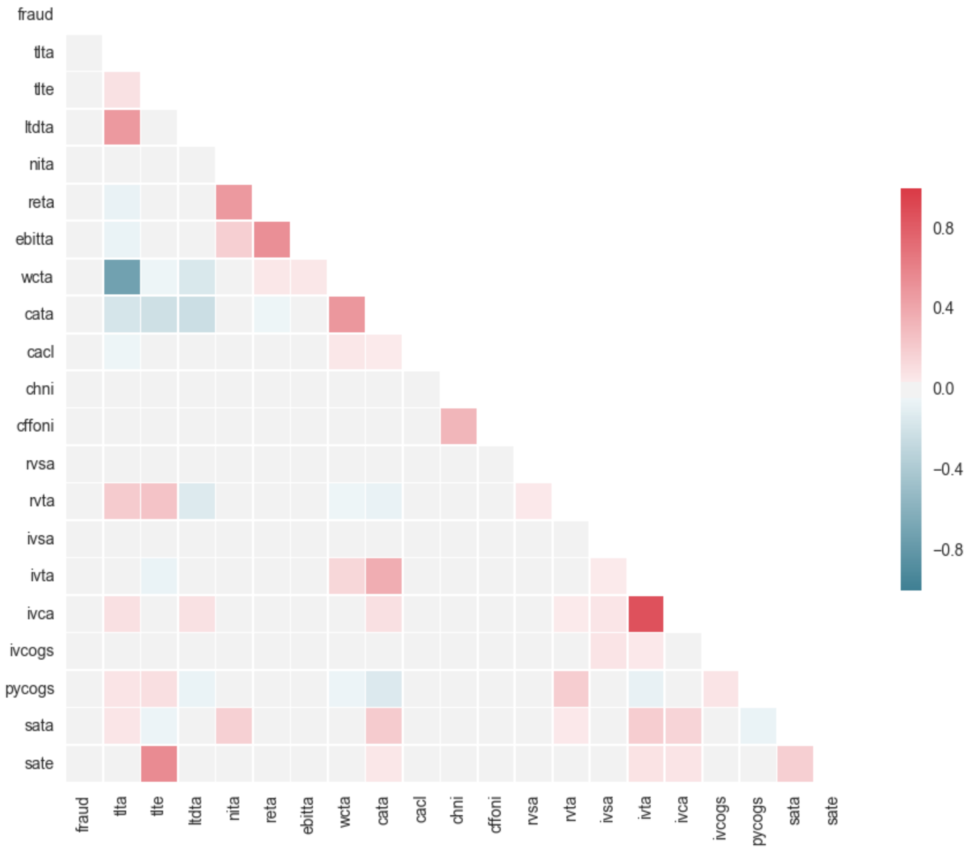
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{4.6}$$

where x and y are the variables of interest, n is the sample size, and $\bar{x}$ and $\bar{y}$ the correspondent sample means.

The aforementioned coefficient suggests a strong positive correlation if the value is close to $+1$, a strong negative correlation if close to -1 and a weak correlation when close to 0. It is important to note that there may be non-linear associations present in the data, but this particular coefficient does not detect such relationships.

The resulting correlation matrix is presented below (Figure 4.2), summarising the correlation coefficients between all financial ratios. A friendly coloured legend is utilised to facilitate visualisation, where red boxes suggest positive linear associations and blue boxes suggest negative linear relations.

FIGURE 4.2: Pearson Correlation Matrix



In order to achieve a better understanding of the magnitude of the linear relation between the financial ratios of interest, a summary of the most relevant Pearson correlation coefficients[3] is presented in Table 4.6.

TABLE 4.6: Most relevant Pearson correlation coefficients

Financial Ratios		Correlation Coefficient
IVTA	IVCA	0.8723
TLTA	WCTA	-0.7215
TLTE	SATE	0.5597
RETA	EBITTA	0.5362
WCTA	CATA	0.4997

Some interesting relationships can be observed from the correlation matrix and summary table. To start with, a strong association between IVTA and IVCA is found. This is not surprising considering both ratios measure the amount of inventory in terms of assets, first considering the total value and then the current amount. As described in Section 3.1, current assets are a component of total assets, so a strong positive correlation is in fact expected between these two items.

In addition, a strong negative relationship between WCTA and TLTA is clearly evident and certainly expected. The rationale behind this relationship is fairly simple: total liabilities is a linear calculation of total assets and total equity (Equation 3.1), and working capital is calculated as the difference between current assets and current liabilities. As current assets are an important component of total assets, then a positive association between WC and TA is expected, as well as a negative association between TL and TA, assuming total equity remain fixed. Consequently, the expectation of a negative relation between WC and TL is reasonably justified.

Moderate positive correlations are also exposed between the following ratios: (i) TLTE and SATE; (ii) RETA and EBITTA; and (iii) WCTA and CATA.

4.5.2 Kendall Correlation

As an alternative approach, Kendall correlation coefficients are calculated next. In this case, the goal is to assess monotonic relationships, that could be linear or not, based-on rank similarity (Kendall, 1955). Monotonic relationships occur when one variable increases as well as the other variable, or when one variable increases and the other one decreases. The increase/decrease of the analysed variables could happen at the same rate, which is the case of linear relations, or in a dissimilar proportion, which is the case of non-linear associations.

[3] Coefficients that score equal or higher than 0.5 in absolute value.

The Kendall correlation, also called Kendall rank correlation coefficient or Kendall's tau, is the non-parametric version of the Pearson correlation. As discussed in the previous section (Chapter 4.3), a non-parametric method makes no assumptions on the distribution of the data, hence the more robust it is when dealing with outliers and data contamination (Sheskin, 2003).

It is said to be a measure of rank correlation in the sense that it calculates the relative position of all observations within one variable (rank position), and then compares them with the ranks obtained within the second variable. If observations from both variables have a similar rank (*concordant* observations), then a high positive correlation will be obtained. Conversely, if ranks are dissimilar (*discordant* observations), then negative correlations are expected.

Kendall correlation can be calculated using the *Tau-A* statistic defined as follows:

$$\tau_A = \frac{n_c - n_d}{n(n-1)/2} \tag{4.7}$$

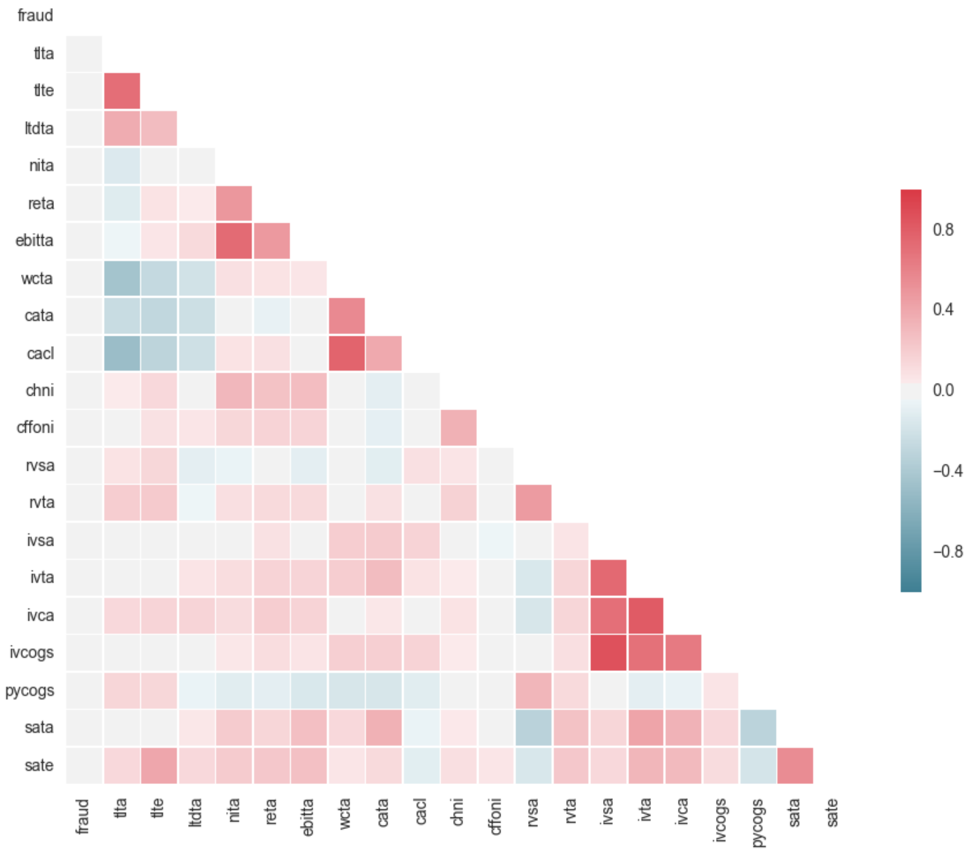
where n_c is the number of concordant pair of observations, n_d is the number of discordant pair of observations, and n is the sample size.

Again, resulting Kendall correlations are summarised in a coloured legend correlation matrix, where intense red boxes indicate positive relationships, i.e.: increases in one variable are associated with increases in the other variable, and intense blue boxes indicate negative associations, that is increases in one variables are aligned with decreases in the other variable. In addition, a summary of most relevant correlations is shown in Table 4.7.

TABLE 4.7: Most relevant Kendall correlation coefficients

Financial Ratios		**Correlation Coefficient**
IVSA	IVCOGS	0.8693
IVTA	IVCA	0.8167
WCTA	CACL	0.7732
IVSA	IVTA	0.7485
NITA	RETA	0.7275
NITA	EBITTA	0.7275
TLTA	TLTE	0.7145
IVSA	IVCA	0.7039
IVCA	IVCOGS	0.6523
WCTA	CATA	0.5684
SATA	SATE	0.5504

FIGURE 4.3: Kendall Correlation Matrix



Some results are in agreement with Pearson correlations, although additional interesting relationships are further exposed.

It can be clearly seen that all inventory-related ratios are strongly positive correlated: IVSA, IVTA, IVCA and IVCOGS. Although this situation is completely expected, it entails an important issue when implementing regression models. If two or more variables are highly correlated then multicollinearity emerges, which means some predictors are redundant. As such, the estimated coefficients of the regression model may be inaccurate, and therefore, not very reliable. A common practice to remedy this problem is to reduce the number of variables in order to keep the most informative ones, procedure that will be implemented at the end of the chapter.

A significant positive association has also been found between CACL and WCTA. This is not surprising considering that WC is actually the subtraction of CA and CL, hence a direct relation between these three financial items results from mathematical construction.

In addition, and as expected, strong positive correlations between ratios related to profitability have been exposed, which includes both NITA and RETA, as well as

NITA and EBITTA. Again, multicollinearity problems may arise when including all these ratios as explanatory variables in a regression model.

A significant positive relation between TLTA and TLTE can also be observed, which is completely expected since total assets, total liabilities and total equity are all connected as a result of Equation 3.1.

Finally, moderate positive correlations have been also exposed between the ratios WCTA and CATA (also found with Pearson's methodology), as well as between SATA and SATE. This latter association makes perfect sense as both ratios are related to sales figures.

4.6 Variable Selection

A comprehensive financial ratio analysis has been conducted to better understand potential predictors of accounting fraud, in particular, which financial information is relevant when detecting fraudulent reports and corporate malpractices.

The goal of variable selection is reducing the number of explanatory variables in a model, as it will make it more concise and fast when classifying a firm as fraud or non-fraud (Baesens, Vlasselaer, and Verbeke, 2015). Then, is suggested to include only meaningful and non-redundant predictors to further achieve satisfactory predictive results.

In what follows, a summary of the removed financial ratios is presented along with an appropriate explanation of the decision made, supported by results obtained from the financial ratio analysis performed before.

- **NITA**: It has been shown that RETA and NITA, as well as RETA and EBITTA are strongly correlated. In order to simplify the analysis and to avoid multicollinearity problems, it is decided to include RETA as the only profitability ratio, mainly because it has shown significant detection power in both parametric and non-parametric methodologies.
- **EBITTA**: It has been determined to exclude EBITTA as explanatory variable using the same argument given before.
- **WCTA**: This ratio is not statistically significant when analysing accounting fraud by industry. In addition, it is strongly correlated with several ratios, such

as TLTA, CACL and CATA, thus it makes no sense to keep it as explanatory variable.

- **CHNI**: This financial ratio shows no significant power as predictor of accounting fraud when implementing both parametric and non-parametric hypothesis testing, hence the decision of remove it from the analysis.
- **RVTA**: A strong association between this ratio and RVSA has been found as a result of the correlation analysis exercise. Previous studies described in Section 2.7 support the usefulness of RVSA as predictor of accounting fraud offences, hence it is decided to keep it as explanatory variable and, in consequence, to remove RVTA.
- **IVCA**: Although it has been shown that this ratio is significant, there is no need to keep it since it is highly correlated with all other inventory-related ratios, including IVSA, IVTA and IVCOGS.
- **SATE**: No significant power was found when implementing the parametric hypothesis testing. In addition, a strong positive correlation with TLTE has been exposed, hence the decision of omitting SATE as explanatory variable.

Finally, 13 financial ratios, out of the original 20, will be selected to continue the analysis of accounting fraud: TLTA, TLTE, LTDTA, RETA, CATA, CACL, CFFONI, RVSA, IVSA, IVTA, IVCOGS, PYCOGS and SATA.

TABLE 4.8: Summary of selected financial ratios and calculation

Selected Ratio	**Calculation**
TLTA	Total Liabilities / Total Assets
TLTE	Total Liabilities / Total Equity
LTDTA	Long-Term Debt / Total Assets
RETA	Retained Earnings / Total Assets
CATA	Current Assets / Total Assets
CACL	Current Assets / Current Liabilities
CFFONI	Cash Flow From Operations / Net Income
RVSA	Accounts Receivable / Total Sales
IVSA	Inventory / Total Sales
IVTA	Inventory / Total Assets
IVCOGS	Inventory / Cost of Good Sold
PYCOGS	Accounts Payable / Cost of Good Sold
SATA	Total Sales / Total Assets

4.7 Summary

In Chapter 4, a complete analysis of financial ratios has been performed. First, the use of ratios as explanatory variables of accounting fraud is justified along with the definition of 20 financial ratios constructed in the basis of financial statements. Then, two commonly used statistical approaches are proposed to assess significant differences between corrupted and genuine reports as well as to identify associations between the considered ratios. Results obtained from hypothesis testing and correlation analysis support the selection of a smaller subset of explanatory variables in both scenarios, first omitting economic domains and later considering them as separate samples.

In the next chapter, a more sophisticated statistical technique will be introduced to assist the process of variable selection, particularly to help choosing the required number of predictors needed for achieving satisfactory detection rates. Again, the analysis will be performed in two stages, first using all observations regardless of sector-specification, and then considering the industry where firms belong to.

Chapter 5

Complete Subset Logistic Regression

Most analytical models implemented to detect fraudulent financial reporting start with numerous variables, out of which only a minority actually contribute to their classification power (Baesens, Vlasselaer, and Verbeke, 2015). Thereby, a question of interest to the public is whether fewer explanatory variables can be used in order to achieve similar accuracy rates as those accomplished when using more predictors.

Classification models can be limited in their performance when a large number of predictors is considered since computational complexity increases as does the risk of overfitting. On the one hand, computational costs are higher as more information has to be analysed, hence the advantage of using only meaningful explanatory variables. On the other hand, considering too many predictors will most definitely make the estimation of the models too complicated, which may lead to an overfitting problem and, subsequently, difficulty of generalising results.

In what follows, an analytical technique is proposed to appropriately tackle the problem of dimensionality reduction and to further justify the number of explanatory variables used when modelling accounting fraud. In particular, an extension of the well-known complete subset regression methodology is implemented using a logistic regression approach as an alternative to the linear model.

Most variable selection techniques, such as best subset regression, stepwise forward regression and stepwise backward regression, aim to select the set of predictors that do the best at meeting some well-defined objective criterion, and then fit a regression model including all selected predictor variables.

On the contrary, what is sought in this chapter is to find the optimal number of variables, which differs to finding the most favourable subset of predictors, as it will give greater flexibility to choose the desired predictor variables based on expert knowledge and prior studies inputs.

This new approach is ultimately assessed using not only the traditional accuracy metrics, but also alternative measurements that are more suitable when dealing with cost-sensitive environments, such as accounting fraud detection.

5.1 Theoretical Background

A very innovative methodology for combining predictions based on complete subset regressions is proposed by Elliot, Gargano, and Timmerman (2013). They suggest this new approach as an alternative to the commonly used forecast methods of ridge regression, model averaging, bagging and the Lasso.

The subset regression approach is adopted as a way to explore how the number of included explanatory variables can be used to trade off the bias and variance of the forecast errors, finding that combinations of subset regressions have the potential to produce more accurate predictions than the conventional techniques previously mentioned, as well as more accurate performance compared to simple linear regression forecasts, which makes this methodology very attractive when reducing the number of variables is desired.

The complete subset regression procedure takes into account multiple forecasting results from the combination of all possible regression models to make a final overall prediction. Accordingly, the number of predictors is fixed, all the regression models considering the same amount of regressors are run, and then a simple average of the forecasts from these regressions is calculated for prediction.

In particular, let be K the number of all possible predictors, out of which K unique simple linear models are directly obtained, as well as $n_{k,K} = K!/((K-k)!k!)$ different k-variate models when considering that $k < K$. A complete subset regression is then defined as the set of all linear regression models for a fixed value of k.

Accordingly, an equal-weighted combination of the forecast from all models within these subsets is calculated as follow:

$$\hat{y} = \frac{1}{K}\sum_{i=1}^{K} X^T \hat{\beta}_i \tag{5.1}$$

where X is the predictors matrix and $\hat{\beta}_i$ are the estimated regression coefficients.

The complete subset regression estimator is then given by:

$$\hat{\beta}_{k,K} = \frac{1}{n_{k,K}} \sum_{i=1}^{n_{k,K}} \hat{\beta}_i \tag{5.2}$$

5.2 Complete Subset Logistic Regression

The proposed methodology described earlier, clearly shows good accuracy performance when applied in a setup akin to the efficient frontier of modern portfolio theory (Elliot, Gargano, and Timmerman, 2013) and when predicting economic affairs such as unemployment levels, GDP growth and inflation rates (Elliot, Gargano, and Timmerman, 2015).

Nevertheless, two major problems arise when modelling the binary fraud target using a linear regression model. On the one hand, the target variable is not normally distributed but rather follows a Bernoulli distribution with only two values, which certainly violates the normality assumption required for the application of a linear model. On the other hand, there is no guarantee that the predicted target will be between 0 and 1, which is an important inconvenience as a binary outcome is desired.

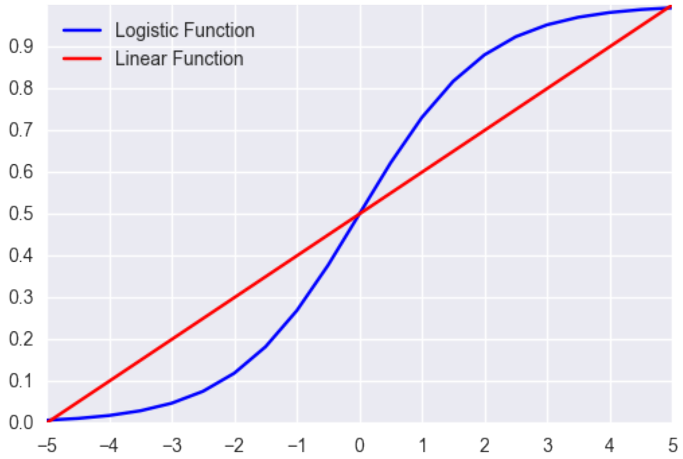
In order to overcome the disadvantages of a linear approach, an alternative methodology is suggested that extends the aforementioned linear regression models to logistic regression models. Thereby, the problem description remains the same except that now a complete subset regression is defined as the set of all logistic regression models for a fixed value of k.

A logistic regression model is defined by the following bounding function:

$$\sigma(z) = \frac{1}{1+\exp(-z)} \tag{5.3}$$

One advantage of the logistic function is that for every possible value of z, the outcome will always be between 0 and 1 (Figure 5.1).

FIGURE 5.1: Bounding Function for Logistic and Linear Regression



Then, extending the original complete subset regression approach, the forecast for all combinations of logistic regression will be:

$$\hat{y} = \sigma(X^T \hat{\beta}) \tag{5.4}$$

Likewise, the complete subset logistic regression estimator will be given by:

$$\hat{\beta}_{k,K} = \frac{1}{n_{k,K}} \sum_{i=1}^{n_{k,K}} \hat{\beta}_i \tag{5.5}$$

It is worth mentioning that implementing the aforementioned logistic version of the complete subset regression approach is not the same as averaging the forecasted logistic functions, i.e.: $\hat{y} = \frac{1}{n_{k,K}} \sum_{i=1}^{n_{k,K}} \sigma_i(X^T \hat{\beta}_i)$. This distinction is critically important as averaging the forecasted estimations is inaccurate when dealing with non-linear functions such as σ_i, hence incorrect calculations of accounting fraud probability are most likely to occur if this formulation is adopted.

To finally decide if an observation is classified as fraudulent or non-fraudulent, then a threshold of 0.5 will be considered. Consequently, the predefined decision rules implemented in this case are the following:

- If $\hat{y} \geq 0.5$, then FRAUD
- If $\hat{y} < 0.5$, then NON-FRAUD

The complete subset logistic regression pseudo code is shown in Algorithm 1.

Algorithm 1 Complete subset logistic regression

```
1: Estimate the null logistic regression model $\mathcal{M}_0$ which contains only the constant.
2: for $k = 1$ to $K$ do
3:     Fit all the $\binom{K}{k}$ possible logistic regression models with exactly $k$ predictors.
4:     Calculate the average of the estimated coefficients.
5:     Fit a logistic regression model using the averaged coefficients and call it $\mathcal{M}_k$.
6: end for
7: Select the best model among $\mathcal{M}_0$, $\mathcal{M}_1$, ..., $\mathcal{M}_K$ according to the aforementioned
   assessment criteria.
```

The proposed complete subset logistic regression approach will be applied next to assist the variable selection procedure started in the previous chapter, as it will be used to determine the optimal number of variables needed to properly model accounting fraud. The analysis will be conducted first using all observations, without taking into account the industry they belong to, and then a more specific evaluation industry by industry will follow.

Before presenting the results obtained by the suggested method, two main issues will be discussed as they are incredibly relevant when dealing with statistical models and accounting fraud offences. In particular, subjects related to imbalance datasets and cost-sensitive environment will be addressed to further achieve more accurate results for both tasks, variable selection and fraud detection.

5.3 Modelling Assessment

An interesting issue related to fraudulent reporting is the difference of misclassification costs. Most studies only seek to maximise overall accuracy without further analysing more suitable assessment measurements.

The cost of misclassification differs when dealing with accounting fraud since a false negative error, which is when a fraudulent observation is classified as non-fraudulent, is usually considered more expensive that a false positive error, which is when a non-fraudulent observation is classified as fraudulent. The reasoning behind this is that a misclassification of a non-fraud firm may cause an important misuse of resources and time, but a misclassification of a fraudulent company may result in incorrect decisions and economic damage.

Accordingly, the overall accuracy rate is no longer sufficient to assess model performance. Other metrics, such as specificity, sensitivity and precision, are now taken into consideration, as well as G-measure, F-measure and AUC, that are calculated using combinations of these metrics. All mentioned indicators are based on the confusion matrix shown in Table 5.1.

TABLE 5.1: Confusion matrix

	Predicted Positives	Predicted Negatives
Real Positives	TP	FN
Real Negatives	FP	TN

Model assessment metrics are described next, including the formula used to calculate them when appropriate.

1. **Overall Accuracy**: it measures the ability to differentiate both fraudulent and genuine observations correctly. It is calculated as the proportion of true positive and true negative cases compared to the total number of observations.

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{5.6}$$

2. **Specificity**: it evaluates the ability to determine non-fraudulent cases correctly. As such, it is computed as the proportion of true negative compared to all legitimate negative observations.

$$specificity = \frac{TN}{TN + FP} \tag{5.7}$$

3. **Sensitivity**: it assesses the capacity to classify fraudulent cases correctly. It is then calculated as the proportion of true positive cases compared to all legitimate positive observations.

$$sensitivity = \frac{TP}{TP + FN} \tag{5.8}$$

4. **Precision**: it measures the proportion of true positive cases compared to all predicted positive observations.

$$precision = \frac{TP}{TP + FP} \tag{5.9}$$

5. **G-Mean**: is the geometric mean of sensitivity and specificity measures. As such, it takes into account the ability of correctly classifying both fraudulent and non-fraudulent observations.

$$G - Mean = \sqrt{sensitivity * specificity} \tag{5.10}$$

6. **F-Measure**: is a metric that integrates both measures of precision and sensitivity

$$F - Measure = \frac{2 * precision * sensitivity}{precision + sensitivity} \tag{5.11}$$

7. **AUC**: The Area Under the Curve (AUC) is a point estimate of the Receiver Operating Characteristic (ROC) curve, which evaluates the diagnostic ability of a binary classifier model as a function of varying a decision threshold. As such, it assesses both true positive and false positive rates considering different threshold settings. The AUC is the probability that the binary classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. As such, AUC is always a positive number range between 0 and 1, so the closer to the unit, the better is the model as it means it is correctly separating instances into the non-fraud and fraud groups. The AUC is computed using the trapezoidal rule, which is a commonly used technique for approximating a definite integral.

Regulatory authorities face critical limitations in terms of human resources, budget support and time constrains, thus a detailed investigation of all records and companies is infeasible or too expensive to undertake. Investigations should concentrate on those firms that are more likely to perpetrate accounting fraud. Therefore, it is preferable to focus on models that correctly classify fraudulent observations rather than non-fraudulent cases.

For this reason, G-measure, F-measure and AUC will be used as model assessment criteria, since they properly capture both false positive and false negative errors, and mitigate the misclassification issue inherent when detecting accounting fraud offences.

5.4 Results

The aforementioned complete subset logistic regression has been adopted to assist in the variable selection task, in particular to evaluate the number of variables required to satisfactory classify fraudulent and genuine financial reports. The implementation of the proposed approach has been done in two scenarios, first using the entire sample and then considering industry specification.

In Chapter 4, a very comprehensive analysis of financial ratios was performed in order to evaluate whether ratio information is useful to detect accounting fraud. Overall, 13 out of the original 20 ratios were found to be significant as explanatory variables of fraudulent reporting. A complementary analysis will be conducted next to further justify the number of predictors needed in different scenarios.

It is worth mentioning, before further interpretation of the results, that all classification accuracy metrics are calculated using out-of-sample data, that is, considering all the data points not belonging to the training sample. Furthermore, the considered model will learn the parameters of a prediction function from a subset of the available data and further tested in a different scenario in order to generalise the results. A standard practice in statistics is to hold out part of the dataset, commonly called *testing set*, and use it later to assess the performance of the model.

Therefore, a stratified 10-fold cross-validation approach is implemented before running the proposed variable selection technique. As such, the studied dataset is divided in 10 folds, each one containing an equal number of fraud and non-fraud cases. For each fold, the model is trained by using the remaining nine folds and then validated by using the hold out fold. At last, model performance is calculated as the average performance of all testing folds (Kirkos, Spathis, and Manolopoulos, 2007).

It can be seen from Table 5.2 that, in general, accuracy metrics show poor classification performance when considering all sampled observations. The optimal number of predictors in this case is somewhere between 2 and 7. The insufficient predictive power may be due to the fact that observations from different industries differ notoriously, hence trying to find a common criminal pattern is a hopeless task.

As implementing only a single model has come at the expense of reduced fraud detection power, then it make sense to build separate models for different industry areas. Accordingly, industry-specific analysis and results are discussed in what follows.

TABLE 5.2: CSLR classification accuracy - All industries

Sample Size: 3,188

k	Accuracy	Specificity	Sensitivity	Precision	G-Mean	F-Measure	AUC
1	0.555	0.278	0.836	0.533	0.482	0.651	0.557
2	0.565	0.344	0.789	0.543	0.521	0.643	0.567
3	0.559	0.365	0.756	0.540	0.525	0.630	0.560
4	0.559	0.373	0.747	0.540	0.528	0.627	0.560
5	0.562	0.386	0.741	0.543	0.535	0.627	0.563
6	0.559	0.388	0.733	0.541	0.533	0.623	0.560
7	0.561	0.394	0.731	0.543	0.537	0.623	0.562
8	0.558	0.392	0.726	0.541	0.534	0.620	0.559
9	0.546	0.386	0.709	0.532	0.523	0.608	0.548
10	0.553	0.388	0.720	0.537	0.529	0.615	0.554
11	0.551	0.388	0.716	0.535	0.527	0.613	0.552
12	0.551	0.396	0.707	0.536	0.529	0.610	0.552
13	0.550	0.400	0.701	0.535	0.530	0.607	0.551

Agriculture, Forestry and Fishing

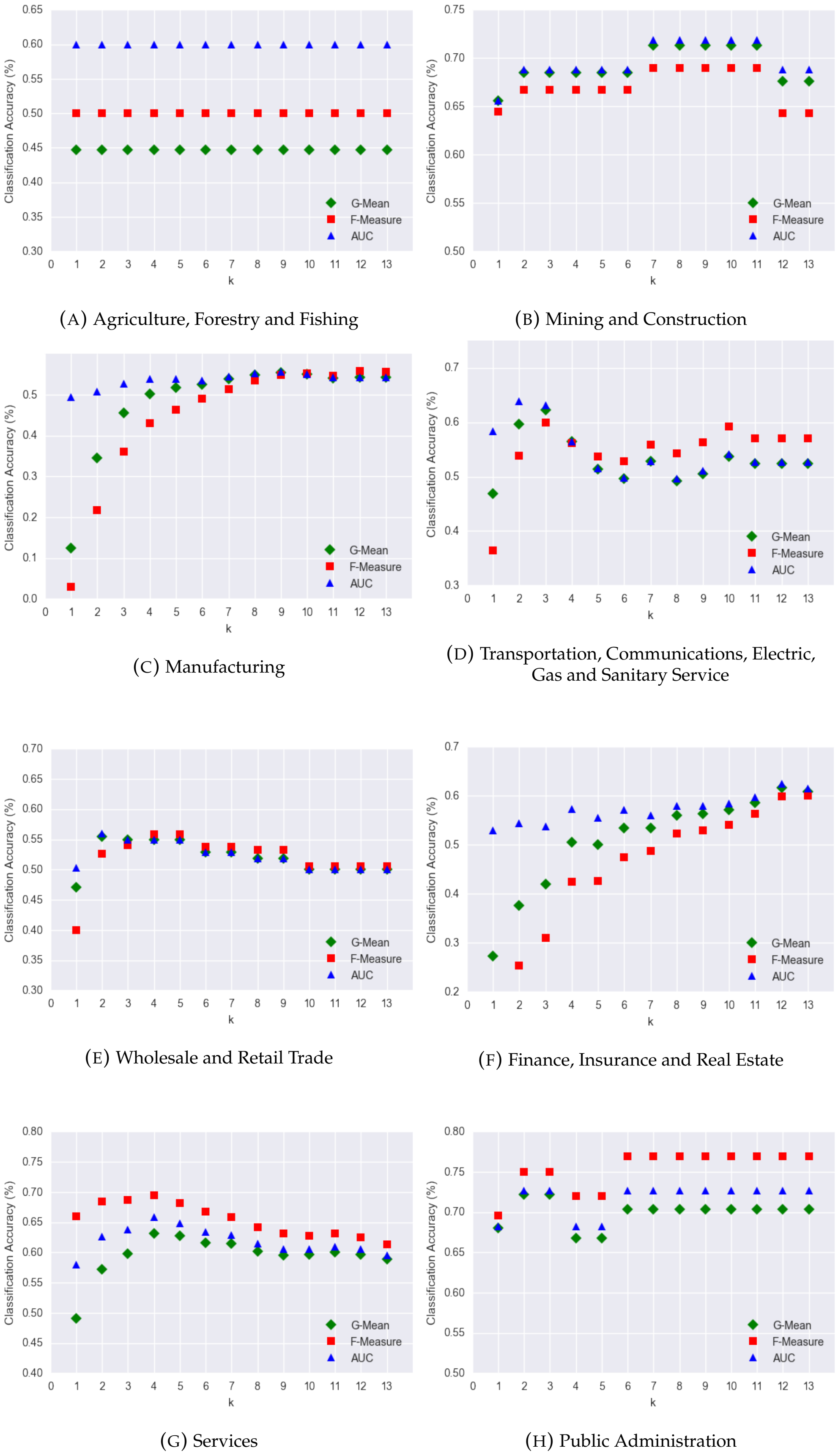
From Table 5.3 and Figure 5.2a, it can be said that the performance of the logistic subset approach does not vary with different values of k in this particular industry. It seems the models are always making the same predictions, regardless of the explanatory variables at issue.

This is probably due to the small size of the studied sample. There are only 22 observations in total, out of which only 11 are fraudulent cases. As such, it is likely that no important differences were found between the groups, hence no classification rules can be made for this industry. In light of this, it is tempting to exclude this industry from the analysis performed as no solid feedback can be drawn. Nonetheless, it is decided to keep observations belonging to this industry as an illustrative exercise where no intentions of generalising results is sought.

Mining and Construction

More meaningful results are obtained in this industry as the sample size increased. It can be observed from Table 5.4 and Figure 5.2b that G-Mean, F-Measure and AUC metrics are larger when considering a number of predictors between 7 and 11.

These results are in accordance with the ones obtained in the ratio analysis performed before, as it was shown that 10 variables were significant when detecting accounting fraud offences. Therefore, there is enough evidence to select them as a reduced set of explanatory variables to be ultimately used for modelling.



(A) Agriculture, Forestry and Fishing

(B) Mining and Construction

(C) Manufacturing

(D) Transportation, Communications, Electric, Gas and Sanitary Service

(E) Wholesale and Retail Trade

(F) Finance, Insurance and Real Estate

(G) Services

(H) Public Administration

FIGURE 5.2: CSLR Classification Accuracy by Industry

TABLE 5.3: CSLR classification accuracy
Industry: Agriculture, Forestry and Fishing

Sample Size: 22							
k	**Accuracy**	**Specificity**	**Sensitivity**	**Precision**	**G-Mean**	**F-Measure**	**AUC**
1	0.429	0.200	1.000	0.333	0.447	0.500	0.600
2	0.429	0.200	1.000	0.333	0.447	0.500	0.600
3	0.429	0.200	1.000	0.333	0.447	0.500	0.600
4	0.429	0.200	1.000	0.333	0.447	0.500	0.600
5	0.429	0.200	1.000	0.333	0.447	0.500	0.600
6	0.429	0.200	1.000	0.333	0.447	0.500	0.600
7	0.429	0.200	1.000	0.333	0.447	0.500	0.600
8	0.429	0.200	1.000	0.333	0.447	0.500	0.600
9	0.429	0.200	1.000	0.333	0.447	0.500	0.600
10	0.429	0.200	1.000	0.333	0.447	0.500	0.600
11	0.429	0.200	1.000	0.333	0.447	0.500	0.600
12	0.429	0.200	1.000	0.333	0.447	0.500	0.600
13	0.429	0.200	1.000	0.333	0.447	0.500	0.600

TABLE 5.4: CSLR classification accuracy
Industry: Mining and Construction

Sample Size: 104							
k	**Accuracy**	**Specificity**	**Sensitivity**	**Precision**	**G-Mean**	**F-Measure**	**AUC**
1	0.656	0.688	0.625	0.667	0.656	0.645	0.656
2	0.688	0.750	0.625	0.714	0.685	0.667	0.688
3	0.688	0.750	0.625	0.714	0.685	0.667	0.688
4	0.688	0.750	0.625	0.714	0.685	0.667	0.688
5	0.688	0.750	0.625	0.714	0.685	0.667	0.688
6	0.688	0.750	0.625	0.714	0.685	0.667	0.688
7	0.719	0.812	0.625	0.769	0.713	0.690	0.719
8	0.719	0.812	0.625	0.769	0.713	0.690	0.719
9	0.719	0.812	0.625	0.769	0.713	0.690	0.719
10	0.719	0.812	0.625	0.769	0.713	0.690	0.719
11	0.719	0.812	0.625	0.769	0.713	0.690	0.719
12	0.688	0.812	0.562	0.750	0.676	0.643	0.688
13	0.688	0.812	0.562	0.750	0.676	0.643	0.688

Manufacturing

Poor classification accuracy is found in the case of manufacturing firms, as it can be seen in Table 5.5 and Figure 5.2c, since values of G-Mean, F-Measure and AUC are fairly small for all subsets.

Nevertheless, slightly better performance is shown when k is between 7 and 10. In addition, ratio analysis results suggested 6 explanatory variables as significant, hence it seems adequate to use them as selected predictors when accounting fraud modelling is required.

TABLE 5.5: CSLR classification accuracy
Industry: Manufacturing

Sample Size: 1,218

k	Accuracy	Specificity	Sensitivity	Precision	G-Mean	F-Measure	AUC
1	0.473	0.971	0.016	0.375	0.124	0.030	0.494
2	0.492	0.880	0.136	0.553	0.346	0.218	0.508
3	0.516	0.794	0.262	0.581	0.456	0.361	0.528
4	0.530	0.737	0.340	0.586	0.501	0.430	0.539
5	0.533	0.691	0.387	0.578	0.518	0.464	0.539
6	0.530	0.634	0.435	0.565	0.525	0.491	0.534
7	0.541	0.623	0.466	0.574	0.539	0.514	0.544
8	0.549	0.606	0.497	0.579	0.549	0.535	0.552
9	0.555	0.594	0.518	0.582	0.555	0.548	0.556
10	0.549	0.566	0.534	0.573	0.550	0.553	0.550
11	0.541	0.554	0.529	0.564	0.541	0.546	0.542
12	0.544	0.531	0.555	0.564	0.543	0.559	0.543
13	0.544	0.537	0.550	0.565	0.543	0.557	0.543

Transportation, Communication, Electric, Gas and Sanitary Service

Ambiguous results are obtained when implementing the proposed methodology in this industry. It can be observed from Table 5.6 and Figure 5.2d that accuracy metrics are larger when the number of predictors is between 1 and 3, and then again when k is 7 and 10, respectively.

Financial ratio analysis showed 5 meaningful variables when detecting accounting fraud offences, that will ultimately be selected for modelling as complete subset approach does not support any other alternative set of potential predictors.

TABLE 5.6: CSLR classification accuracy
Industry: Transportation, Communication, Electric, Gas and Sanitary Service

Sample Size: 212

k	Accuracy	Specificity	Sensitivity	Precision	G-Mean	F-Measure	AUC
1	0.562	0.933	0.235	0.800	0.469	0.364	0.584
2	0.625	0.867	0.412	0.778	0.597	0.538	0.639
3	0.625	0.733	0.529	0.692	0.623	0.600	0.631
4	0.562	0.600	0.529	0.600	0.564	0.562	0.565
5	0.516	0.500	0.529	0.545	0.514	0.537	0.515
6	0.500	0.467	0.529	0.529	0.497	0.529	0.498
7	0.531	0.500	0.559	0.559	0.529	0.559	0.529
8	0.500	0.433	0.559	0.528	0.492	0.543	0.496
9	0.516	0.433	0.588	0.541	0.505	0.563	0.511
10	0.547	0.467	0.618	0.568	0.537	0.592	0.542
11	0.531	0.467	0.588	0.556	0.524	0.571	0.527
12	0.531	0.467	0.588	0.556	0.524	0.571	0.527
13	0.531	0.467	0.588	0.556	0.524	0.571	0.527

Wholesale Trade and Retail Trade

Poor classification accuracy is found in the case of trading firms as values of G-Mean, F-Measure and AUC are fairly small for all subsets. Nevertheless, it can be observed from Table 5.7 and Figure 5.2e that slightly better performance is achieved when k is between 2 and 4, which is in agreement with the results obtained in the ratio analysis, as it suggested 3 explanatory variables as significant. Consequently, it makes sense to use them as selected predictors when accounting fraud modelling is performed.

TABLE 5.7: CSLR classification accuracy
Industry: Wholesale Trade and Retail Trade

Sample Size: 338

k	Accuracy	Specificity	Sensitivity	Precision	G-Mean	F-Measure	AUC
1	0.500	0.680	0.327	0.515	0.471	0.400	0.503
2	0.559	0.640	0.481	0.581	0.555	0.526	0.560
3	0.549	0.580	0.519	0.562	0.549	0.540	0.550
4	0.549	0.540	0.558	0.558	0.549	0.558	0.549
5	0.549	0.540	0.558	0.558	0.549	0.558	0.549
6	0.529	0.520	0.538	0.538	0.529	0.538	0.529
7	0.529	0.520	0.538	0.538	0.529	0.538	0.529
8	0.520	0.500	0.538	0.528	0.519	0.533	0.519
9	0.520	0.500	0.538	0.528	0.519	0.533	0.519
10	0.500	0.500	0.500	0.510	0.500	0.505	0.500
11	0.500	0.500	0.500	0.510	0.500	0.505	0.500
12	0.500	0.500	0.500	0.510	0.500	0.505	0.500
13	0.500	0.500	0.500	0.510	0.500	0.505	0.500

Finance, Insurance and Real Estate

Interesting results are observed in the financial industry, as better performance is obtained when more variables are included in the analysis. As it can be seen in Table 5.8 and Figure 5.2f, G-Mean, F-Measure and AUC metrics are larger when considering a number of predictors larger or equal than 8.

These results are in accordance with the outcome obtained in the ratio analysis where exactly eight variables were significant. Hence, there is enough evidence to select them as a reduced set of explanatory variables to be ultimately used when implementing a fraud detection mechanism.

TABLE 5.8: CSLR classification accuracy
Industry: Finance, Insurance and Real Estate

Sample Size: 472

k	Accuracy	Specificity	Sensitivity	Precision	G-Mean	F-Measure	AUC
1	0.479	0.984	0.076	0.857	0.273	0.140	0.530
2	0.500	0.937	0.152	0.750	0.377	0.253	0.544
3	0.500	0.873	0.203	0.667	0.420	0.311	0.538
4	0.542	0.841	0.304	0.706	0.506	0.425	0.573
5	0.528	0.794	0.316	0.658	0.501	0.427	0.555
6	0.549	0.778	0.367	0.674	0.534	0.475	0.572
7	0.542	0.730	0.392	0.646	0.535	0.488	0.561
8	0.563	0.730	0.430	0.667	0.561	0.523	0.580
9	0.563	0.714	0.443	0.660	0.563	0.530	0.579
10	0.570	0.714	0.456	0.667	0.571	0.541	0.585
11	0.585	0.714	0.481	0.679	0.586	0.563	0.598
12	0.613	0.730	0.519	0.707	0.616	0.599	0.625
13	0.606	0.698	0.532	0.689	0.609	0.600	0.615

Services

Table 5.9 and Figure 5.2g summarise the results obtained when adopting a complete subset approach in the service industry. It can be observed better performance when k is between 3 and 6, as accuracy metrics are larger in this range of subsets.

One more time, these results are in accordance with the ratio analysis performed in Chapter 4.1, as 6 variables were found to be significant as predictors of accounting fraud offences. Therefore, it makes sense to use them as explanatory variables for detecting fraudulent financial reports.

TABLE 5.9: CSLR classification accuracy
Industry: Services

Sample Size: 750

k	Accuracy	Specificity	Sensitivity	Precision	G-Mean	F-Measure	AUC
1	0.564	0.271	0.888	0.525	0.491	0.660	0.580
2	0.613	0.373	0.879	0.560	0.572	0.684	0.626
3	0.627	0.415	0.860	0.571	0.598	0.687	0.638
4	0.649	0.475	0.841	0.592	0.632	0.695	0.658
5	0.640	0.483	0.813	0.588	0.627	0.682	0.648
6	0.627	0.483	0.785	0.579	0.616	0.667	0.634
7	0.622	0.492	0.766	0.577	0.614	0.659	0.629
8	0.609	0.492	0.738	0.568	0.602	0.642	0.615
9	0.600	0.492	0.720	0.562	0.595	0.631	0.606
10	0.600	0.500	0.710	0.563	0.596	0.628	0.605
11	0.604	0.508	0.710	0.567	0.601	0.631	0.609
12	0.600	0.508	0.701	0.564	0.597	0.625	0.605
13	0.591	0.508	0.682	0.557	0.589	0.613	0.595

Public Administration

Finally, an exceptional performance of the proposed methodology can be seen for public administration institutions as accuracy measures are fairly large for all values of k. In particular, better results are obtained when the number of predictors is 6 or more, similar to what was found in the financial ratio analysis. In consequence, 8 ratios will be used to further model accounting fraud within this economic sector.

TABLE 5.10: CSLR classification accuracy
Industry: Public Administration

Sample Size: 72							
k	**Accuracy**	**Specificity**	**Sensitivity**	**Precision**	**G-Mean**	**F-Measure**	**AUC**
1	0.682	0.636	0.727	0.667	0.680	0.696	0.682
2	0.727	0.636	0.818	0.692	0.722	0.750	0.727
3	0.727	0.636	0.818	0.692	0.722	0.750	0.727
4	0.682	0.545	0.818	0.643	0.668	0.720	0.682
5	0.682	0.545	0.818	0.643	0.668	0.720	0.682
6	0.727	0.545	0.909	0.667	0.704	0.769	0.727
7	0.727	0.545	0.909	0.667	0.704	0.769	0.727
8	0.727	0.545	0.909	0.667	0.704	0.769	0.727
9	0.727	0.545	0.909	0.667	0.704	0.769	0.727
10	0.727	0.545	0.909	0.667	0.704	0.769	0.727
11	0.727	0.545	0.909	0.667	0.704	0.769	0.727
12	0.727	0.545	0.909	0.667	0.704	0.769	0.727
13	0.727	0.545	0.909	0.667	0.704	0.769	0.727

Lastly, a summary table is provided below (Table 5.11) including financial ratios that have been selected for each industry.

Interesting differences between sectors emerge from the previously performed analysis as some ratios are significant or not depending on the industry the company belongs to.

On one hand, inventory and retained earnings are relevant predictors in the industries of transportation, communication, electric gas and sanitary service, wholesale trade and retail trade, and services. This may be due to the fact that inventory volumes and retained earning are easily falsified within the aforementioned sectors.

On the other hand, manufacturing companies may be tempted to modify items related to liabilities as well as current assets, while finance, insurance and real estate firms manipulate liabilities and cash flow from operation figures.

5.5 Summary

In Chapter 5, a simple yet novel approach is expanded from the well-known complete subset regression approach to ultimately assist the selection of significant explanatory variables performed in the previous chapter. It is interesting to see that results from the proposed methodology strongly support the number of selected ratios as predictors of accounting fraud in two scenarios, first omitting industry features and then using domain-specific samples.

In the next chapter, several machine learning models will be implemented in order to achieve satisfactory detection rates of accounting fraud offences. In this regard, basic and more complex predictive models are described and adopted for determining the likelihood of accounting fraud occurrence along with interpretation of overall and industry-specific results.

TABLE 5.11: Summary of selected financial ratios by industry domain

Mining and Construction	10	TLTA TLTE LTDTA RETA CACL RVSA IVTA IVCOGS PYCOGS SATA
Manufacturing	6	TLTA TLTE RETA CATA CACL RVSA
Transportation, Communications, Electric, Gas and Sanitary Service	5	RETA IVSA IVTA PYCOGS SATA
Wholesale Trade and Retail Trade	3	RETA CATA IVSA
Finance, Insurance and Real Estate	8	TLTA TLTE LTDTA RETA CFFONI IVCOGS PYCOGS SATA
Services	6	RETA CACL IVSA IVCOGS PYCOGS SATA
Public Administration	8	LTDTA RETA CATA CACL IVSA IVTA IVCOGS SATA

Chapter 6

Accounting Fraud Modelling

6.1 Statistical Modelling

Implementation and assessment of analytical models has already been accomplished in the previous chapter to assess the selection of significant explanatory variables. In particular, an extension of the complete subset linear regression has been adopted using a logistic approach instead. More accurate results have been obtained as more advanced methodologies have been implemented, which certainly motivates the use of statistical models for accounting fraud detection.

In what follows, several machine learning methods will be properly described, implemented and further interpreted in order to achieve satisfactory detection rates. Moreover, a breakdown of the analysis is adopted considering the industry in which firms belong to, by virtue of exposing domain-specific fraudulent behaviours.

6.2 Machine Learning Methods

The binary outcome model is considered to be the foundational scheme for detecting accounting fraud since the aim is to classify future observations into only two possible values: fraud or non-fraud.

Accordingly, this study assesses the effectiveness of several machine learning models in the identification of fraudulent reporting. First, Discriminant Analysis and Logistic Regression are employed as benchmark framework followed by the implementation of more advanced but easy-to-interpret algorithms such as AdaBoost, Decision Trees, Boosted Trees and Random Forests.

The motivation for using boosting techniques and tree-based methods is supported in part by the poor detection accuracy of basic models and in part by the excessive complexity of more sophisticated approaches, such as neural networks and support vector machines.

In order to achieve a consistent notation throughout the chapter, the following conventions are used for mathematical equations:

- A superscript T denotes the transpose of a matrix or vector.
- $Y = 1$: fraudulent observation.
- $Y = 0$: non-fraudulent observation.
- $P(Y = 1 \mid X)$: posterior probability of fraud.
- $P(Y = 0 \mid X)$: posterior probability of non-fraud.

It is worth noting that given there are only two possible outcomes, then it holds that:

$$P(Y = 0 \mid X) = 1 - P(Y = 1 \mid X) \tag{6.1}$$

The models were employed as implemented in the Scikit-Learn library (Pedregosa et al., 2011) and an exhaustive explanation of each algorithm is given below.

6.2.1 Discriminant Analysis

Discriminant analysis is a supervised method used in statistics to address classification problems and to make predictions of a categorical dependent variable. The main idea is to classify an observation into one of the predefined classes using a combination of one or more continuous independent variables in order to generate a discriminant function which best differentiate between the groups.

Subsequently, a decision boundary is generated by fitting class conditional densities $P(X \mid Y)$ to the data using Bayes' rule:

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)} = \frac{P(X \mid Y)P(Y)}{\sum_y P(X \mid Y = y)P(Y = y)} \tag{6.2}$$

The appropriate class is selected which maximises these conditional probabilities. In the case of accounting fraud, only two classes are of interest; therefore:

$$P(Y = 0 \mid X) = \frac{P(X \mid Y = 0)P(Y = 0)}{P(X \mid Y = 0)P(Y = 0) + P(X \mid Y = 1)P(Y = 1)} \tag{6.3}$$

$$P(Y = 1 \mid X) = \frac{P(X \mid Y = 1)P(Y = 1)}{P(X \mid Y = 0)P(Y = 0) + P(X \mid Y = 1)P(Y = 1)} \tag{6.4}$$

The optimisation task is ultimately achieved using the training data to estimate class priors, both $P(Y = 0)$ and $P(Y = 1)$, class means and the covariance matrices. In particular, class priors are estimated as the proportion of instances in each class, that is, number of fraudulent (or non-fraudulent) observation divided by the total number of observations. Class means are estimated using the empirical sample class means. Similarly, covariance matrices are estimated using the empirical sample class covariance matrices.

In accordance with the aforementioned, the following assumptions are made:

1. Predictors are all statistically independent.
2. $P(X \mid Y)$ follows a multivariate Gaussian distribution, with a class-specific mean and covariance matrix.

Different assumptions associated with the covariance matrix will lead to different decision boundaries, one defined by a linear combination of the predictors and another one by a quadratic form.

In both cases, however, the predicted class will be determined using a classification threshold of 0.5. As such, if the estimated probability of fraud occurrence ($P(Y = 1)$) is equal or higher than 0.5, then the observation will be classified as fraudulent. On the contrary, if $P(Y = 1)$ is lower than 0.5, or equivalently $P(Y = 0) \geq 0.5$, then the observation will be classified as non-fraudulent.

Linear Discriminant Analysis

In the particular case of Linear Discriminant Analysis (LDA), a multivariate normal distribution of the predictors is presumed with a distinct mean for each class and a covariance matrix that is common to all classes. For AF detection, this means that both fraud and non-fraud classes share the same covariance matrix $\Sigma_0 = \Sigma_1 = \Sigma$.

The advantage of a common covariance matrix is that it simplifies the problem by reducing the computational cost of estimating a large number of parameters when the number of predictors is relatively large. Taking this into consideration, then it is true that:

$$P(X \mid Y=0) = \frac{1}{(2\pi)^n \left|\Sigma\right|^{1/2}} exp\left(-\frac{1}{2}(X-\mu_0)^T\Sigma^{-1}(X-\mu_0)\right) \tag{6.5}$$

$$P(X \mid Y=1) = \frac{1}{(2\pi)^n \left|\Sigma\right|^{1/2}} exp\left(-\frac{1}{2}(X-\mu_1)^T\Sigma^{-1}(X-\mu_1)\right) \tag{6.6}$$

Quadratic Discriminant Analysis

Furthermore, Quadratic Discriminant Analysis (QDA) provides a similar approach yet now it is assumed that the covariance matrix is class-specific, i.e.: $X \sim N(\mu_k, \Sigma_k)$ for the kth class. Therefore:

$$P(X \mid Y=0) = \frac{1}{(2\pi)^n \left|\Sigma_0\right|^{1/2}} exp\left(-\frac{1}{2}(X-\mu_0)^T\Sigma_0^{-1}(X-\mu_0)\right) \tag{6.7}$$

$$P(X \mid Y=1) = \frac{1}{(2\pi)^n \left|\Sigma_1\right|^{1/2}} exp\left(-\frac{1}{2}(X-\mu_1)^T\Sigma_1^{-1}(X-\mu_1)\right) \tag{6.8}$$

6.2.2 Logistic Regression

Logistic Regression (LR) models are commonly used for performing binary classification. As described in Chapter 5.2, the goal is to fit a regression model that estimates the accounting fraud likelihood applying a logistic function that is linear in its argument:

$$\sigma(Z) = \frac{1}{1+exp(-Z)} \tag{6.9}$$

In order to obtain the best classification possible, the posterior probability of belonging to one of both categories is calculated by maximising the likelihood function. Likewise, let $P(Y=1 \mid X)$ be the posterior probability of fraud (Bishop, 2006), then:

$$P(Y=1 \mid X) = y(X) = \sigma(w^T X) \tag{6.10}$$

For a dataset $\{x_n, t_n\}$, where $t_n \in$ {0,1} and $n = 1, ..., N$, the likelihood of any specific outcome is given by:

$$P(t \mid w) = \sum_n y_n^{t_n}\{1 - y_n\}^{1-t_n} \tag{6.11}$$

where $t = (t_1, ..., t_N)^T$ and $y_n = P(Y = 1 \mid x_n)$.

As mentioned before, the maximum likelihood estimates of w are obtain by minimising the cross-entropy error function defined by the negative logarithm of the likelihood and then taking its gradient with respect to w:

$$E(w) = -ln\{P(t \mid w)\} = -\sum_n \{t_n ln(y_n) + (1 - t_n) ln(1 - y_n)\} \tag{6.12}$$

$$\bigtriangledown E(w) = \sum_n (y_n - t_n) x_n \tag{6.13}$$

To finally decide if an observation is classified as fraudulent or non-fraudulent, then a threshold of 0.5 will be considered. Consequently, if $P(Y = 1 \mid X)$ is estimated to be equal or greater than 0.5, then the observation will be classified as fraudulent. Otherwise, it will be classified as non-fraudulent.

In this section, a logistic regression methodology will be implemented using all explanatory variables of interest rather than the subset approach previously adopted in Chapter 5.

More advanced statistical models will be introduced next. As mentioned in Section 2.7, several complex machine learning methods have been developed in previous studies to detect fraudulent accounting records, such as Bayesian networks, support vector machines and hybrid algorithms. The achieved performance of all these techniques are quite superior to more basic methods, but the cost of this improvement is somehow high when taking into account the considerable drawbacks that these methods entail, including important computational costs and overfitting proneness, as well as struggling when interpreting results.

Consequently, decision trees and boosting techniques are suggested to be implemented as their advantages can be very useful when detecting accounting fraud. More details about the proposed methods will be discuss in what follows.

6.2.3 AdaBoost

Adaptive boosting, widely known as AdaBoost (AB), is a machine learning technique used for classification and regression problems that combines multiple 'weak learner' classifiers in order to produce a better boosted classifier. In this context, a weak learner is a function that is only weakly correlated with the response.

The basic idea is to weight observations w_n by how easy or difficult they are to categorise, giving more importance to those that are harder to predict in order to learn from them and further construct better subsequent classifiers.

Accordingly, each individual classifier generates an output $G_m(X)$, $m = 1, ..., M$, for every observation n of the training set. Then, these classifiers are trained on a weighted form using α_m as classifier coefficients. As mentioned before, misclassified instances will be given greater weight when used to train the subsequent classifier (Bishop, 2006).

The goal is to minimise a weighted error function err_m in every iteration m taking into account the information and performance of previous classifiers. Ultimately and after the last iteration M, a final boost classifier $G(X)$ is constructed as an additive combination of all trained weak learner classifiers $G_m(X)$:

$$G(X) = sign[\sum_m \alpha_m G_m(X)] \tag{6.14}$$

In this case, a classification threshold of 0.5 has been adopted. As such, an observation will be classified as fraudulent when $G(X)$ is equal or greater than 0.5, and classified as non-fraudulent when $G(X)$ is lower than 0.5.

The AdaBoost pseudo code is shown in Algorithm 2[1].

[1] Scharth, M. (2017). Statistical Learning and Data Mining, Module 15 [PowerPoint presentation]. Discipline of Business Analytics, The University of Sydney Business School, QBUS6810.

Algorithm 2 AdaBoost

```
Initialise the observation weights $w_n = 1/N, n = 1, ..., N$.
for $m = 1$ to $M$ do
    Fit a classifier $G_m(X)$ to the training data using weights $w_n$.
    Compute the weighted error rate.
        $err_m = \frac{\sum_n w_n I(y_n \neq G - m(x_n))}{\sum_n w_n}$
    Compute $\alpha_m = log((1 - err_m)/err_m)$.
    Update the weights,
        $w_n \leftarrow w_n exp[\alpha_m I(y_n \neq G_m(x_n)]$
end for
Output the classification $G(X) = \text{sign}[\sum_m \alpha_m G_m(X)]$.
```

6.2.4 Decision Trees

Decision Trees (DT) are a non-parametric supervised learning method that classify observations based on the values of one or more predictors. The advantage of decision trees lies in the straightforward extraction of if-then classification rules easily replicable by auditors and regulatory authorities. Also, no assumptions on the structure of the data is needed, which is very convenient in this case considering the asymmetrical distribution of some explanatory variables.

The structure of a DT consists of nodes representing a test on a particular attribute and branches representing an outcome of the test. The idea is to divide observations into mutually exclusive classes in order to build the smallest set of rules that is consistent with the training data. To identify the attribute that best separates the sample, information gain and entropy reduction are used as estimation criteria.

There are several tree algorithms, such as ID3, C4.5, C5.0 and CART, among others. The chosen method used in this study is the Classification and Regression Trees (CART) characterised by the construction of binary trees based-on feature and threshold selection that provide the largest information gain in each node. This algorithm recursively partitions the space in order to minimise the error or impurity of each node, resulting in terminal nodes that represent homogeneous groups that differ substantially from the others.

Accordingly, let the information at node m be Q, then the binary partition of the data is defined by a candidate split θ that divides the space into two subsets: $Q_{left}(\theta)$ and $Q_{right}(\theta)$.

The error at node m is calculated using an impurity function H evaluated in both partitions, that later is minimised in order to estimate the parameters.

$$G(Q, \theta) = \frac{n_{left}}{N_m} H(Q_{left}(\theta)) + \frac{n_{right}}{N_m} H(Q_{right}(\theta)) \tag{6.15}$$

$$\theta^* = argmin_\theta G(Q, \theta) \tag{6.16}$$

The impurity function implemented in this study corresponds to the Gini function:

$$H(X_m) = \sum_k p_{mk}(1 - p_{mk}) \tag{6.17}$$

where p_{mk} is the proportion of class k observations in node m.

It is worth noting that the partitions of the predictor space are based on a greedy algorithm called *recursive binary splitting*. The technique is greedy because at the best split is made at each step of the tree-building process without taking into account the consequences further down the tree. Consequently, in some cases very complex trees are generated as result of this approach, what is commonly known in statistical jargon as *overfitting*. However, a couple of mechanisms can be used in order to avoid this situation, such as setting the minimum number of required observations at a leaf node or setting the maximum depth of the tree.

The tree size is therefore a tuning parameter determining the complexity of the model and it should be selected adaptively from the data. As such, the maximum number of node splits in the current study is settled as 5, optimal valued obtained by cross validation.

Decision trees are remarkably superior than the first two methods mentioned before - logistic regression and discriminant analysis - considering how easy they are to explain, implement and visualise. Unfortunately, they show some drawbacks that should be mentioned, such as their inherent instability that emerges when little changes in the data cause a large change in the structure of the estimated tree, as well as the lower predictive accuracy when compared to more advanced techniques.

Decision trees can be used as the basic component of powerful prediction methods. Therefore, two additional models that employ decision trees as their foundation, will be introduce below.

6.2.5 Boosted Trees

Similar to the AdaBoost approach, Boosted Trees (BT) is an ensemble of weak learners but now in the explicit form of fixed size decision trees as base classifiers.

Accordingly, an iterative process takes place in order to fit a decision tree output $h_m(X)$ in every iteration m to improve the previous model $F_m(X)$ by constructing a new model that adds this new information:

$$F_{m+1}(X) = F_m(X) + h_m(X) \tag{6.18}$$

The main idea is to minimise an error function defined by the difference between the old model $F_m(X)$ and the new one $F_{m+1}(X)$, what is called the *residual*, through a gradient boosting algorithm that is much like the gradient descent method used in the logistic regression approach.

In this case, a classification threshold of 0.5 has also been adopted. In this regard, an observation will be classified as fraudulent when $F_m(X)$ is equal or greater than 0.5, and classified as non-fraudulent when $F_m(X)$ is lower than 0.5.

Same as in the decision tree methodology and for consistency, the maximum depth of the fitted trees is established to be 5.

6.2.6 Random Forests

A further enhancement of boosted trees is provided by the Random Forests (RF) approach, one of the most popular bagging techniques. Bootstrap aggregation, or bagging, averages many noisy but approximately unbiased models, which results in a reduction of the variance.

The idea is to fit a classification model to the training data $\mathcal{D}$ to obtain the prediction $\hat{f}(X)$. Bagging averages this prediction over a collection of bootstrap samples[2]. For each bootstrap sample $\mathcal{D}_b^*$, $b = 1, ..., B$, the selected classification model is fitted to obtain a prediction $\hat{f}_b^*(X)$. The bagged classifier selects the class (fraud or non-fraud) with the most "votes" from the B classifiers:

$$\hat{y}_{bag}(X) = \arg\max_c \sum_b I(\hat{y}_b^*(X) = c) \tag{6.19}$$

Decision trees are ideal candidates for bagging as they capture complex interactions structures in the data, which leads to relatively low biased but high variance. Consequently, classification trees are adopted next for bagging to further construct Random Forests.

Random Forests improve over bagging by adding an adjustment that helps decorrelate the trees. In this context, instead of using all predictors, RF only selects a random subset of the features as split candidates in each step. The rationale behind this methodology is that when establishing a fewer and fixed number of predictors, then more variation in the structure of the model is allowed, which diminishes the correlation between the resulting trees. Interestingly, this new condition makes the average of the fitted trees less variable and therefore more reliable (James et al., 2013).

In building a random forest, k independent variables out of all possible predictors are randomly selected at each node, and later the best split on the considered variables is found. As a last step, all trees are averaged to obtain a final prediction.

The Random Forests pseudo code is shown in Algorithm 3[3].

[2] In statistics, bootstrapping is any test or metric that relies on random sampling with replacement.

[3] Scharth, M. (2017). Statistical Learning and Data Mining, Module 13 [PowerPoint presentation]. Discipline of Business Analytics, The University of Sydney Business School, QBUS6810.

Algorithm 3 Random Forest

```
for $b = 1$ to $B$ do
    Sample $N$ observations with replacement from the training data $\mathcal{D}$ to obtain the bootstrap sample $\mathcal{D}_b^*$.
    Grow a random forest tree $T_b$ to $\mathcal{D}_b^*$ by repeating the following steps for each terminal node of the tree, until the minimum node size is reached:
    (i) Select $k$ variables at random from the $K$ variables.
    (ii) Pick the best variable and split point among the $k$ candidates.
    (iii) Split the node into two daughter nodes.
end for
Output the ensemble of trees $\{T_b\}_1^B$.
```

In order to be consistent with the previous methodologies, the maximum depth of the estimated trees is established to be 5.

6.3 Results

The main objective of machine learning is to correctly make predictions on data. For this end, and to avoid overfitting, a model should learn the parameters of a prediction function from a subset of the available data and be tested in a different scenario in order to generalise the results. A standard practice in statistics is to hold out part of the dataset, commonly called *testing set*, and use it later to assess the performance of the model.

Therefore, a stratified 10-fold cross-validation approach is implemented before running the proposed techniques. As such, the studied dataset is divided in 10 folds, each one containing an equal number of fraud and non-fraud cases. For each fold, the models are trained by using the remaining nine folds and then validated by using the hold out fold. At last, models performance is calculated as the average performance of all testing folds (Kirkos, Spathis, and Manolopoulos, 2007).

It is worth recalling that explanatory variables used to train and test the proposed models are the thirteen financial ratios selected in Chapter 4 and later validated in Chapter 5: TLTA, TLTE, LTDTA, RETA, CATA, CACL, CFFONI, RVSA, IVSA, IVTA, IVCOGS, PYCOGS and SATA.

In order to accurately evaluate the predictive power of the proposed machine learning methods, all assessment metrics defined in Section 5.3 are calculated using the testing data exclusively, that is, using the out-of-sample data.

Table 6.1 reports the results of the proposed models considering observations from all sectors indistinctly of the industries they belong.

TABLE 6.1: General Prediction Accuracy

Sample size: 3,188	Accuracy	Specificity	Sensitivity	Precision	G-Mean	F-Measure	AUC
LDA	0.551	0.490	0.607	0.562	0.545	0.583	0.549
QDA	0.555	0.163	0.919	0.542	0.387	0.682	0.541
LR	0.551	0.449	0.645	0.557	0.538	0.598	0.547
AB	0.577	0.566	0.587	0.593	0.576	0.590	0.576
DT	0.565	0.295	0.817	0.555	0.491	0.661	0.556
BT	0.608	0.588	0.627	0.621	0.607	0.624	0.607
RF	0.554	0.482	0.621	0.563	0.547	0.591	0.551

It can be observed that overall, models are not showing exceptional performance. Average accuracy rates indicate that the proposed models are at least slightly better than random chance, that is, classifying 50% of observations in each group. Furthermore, sensitivity rates are considerably higher than specificity values, which implies that the tested algorithms tend to classify more instances as fraudulent than as non-fraudulent.

Particularly, the largest sensitivity rate is performed by QDA which also shows the lowest specificity rate, indicating that its predictions are excessively favouring the positive cases. Similar situation is observed when implementing a decision tree approach.

Better performance can be seen in the case of boosted trees as this technique shows the highest values of G-Mean and AUC, as well as a decent rate of F-Measure, suggesting that correct classification of non-fraud and fraud cases is performed evenly when using this approach. In particular, a sensitivity rate of 0.627 is achieved by this model, which indicates a correct classification of 62.7% of fraudulent companies.

As discovered in previous sections, different accounting tricks are prone to be adopted when fraud is being committed within different industries. In consequence, implementing machine learning models without distinguishing in which sector each observation belongs to, will most likely lead to poor classification performance. As such, a breakdown of the analysis is performed by industry, considering two scenarios, first using the complete set of predictor ratios, and then using the proposed reduced set of ratios for each industry, meticulously detailed in the previous chapter.

Results are reported in what follows, including a detailed analysis of models performance and selected variables, this time industry by industry.

Agriculture, Forestry and Fishing

Good classification performance is achieved in this industry, as shown in Table 6.2. Dissimilar results are obtained when considering different number of predictors whatsoever, probably due to the small size of the sample at issue.

Figure 6.1 expose better results for logistic regression when using all financial ratios, as well as extraordinary performance of QDA and AdaBoost when only the reduced set of ratios is considered, as their specificity and sensitivity metrics notably exceed other models. In both cases, 75% of non-fraudulent cases are correctly identified, as well as 100% of fraudulent cases. Special care must be taken when generalising these results, as a fairly small sample is being considered.

TABLE 6.2: Classification Accuracy Industry Specific
SIC 1: Agriculture, Forestry and Fishing

Sample size: 22	Accuracy	Specificity	Sensitivity	Precision	G-Mean	F-Measure	AUC
Complete set of financial ratios $K = 13$							
LDA	0.429	0.500	0.333	0.333	0.408	0.333	0.417
QDA	0.571	0.250	1.000	0.500	0.500	0.667	0.625
LR	0.857	0.750	1.000	0.750	0.866	0.857	0.875
AB	0.571	0.750	0.333	0.500	0.500	0.400	0.542
DT	0.714	0.750	0.667	0.667	0.707	0.667	0.708
BT	0.857	0.750	1.000	0.750	0.866	0.857	0.875
RF	0.429	0.250	0.667	0.400	0.408	0.500	0.458
Reduced set of financial ratios $k = 4$							
LDA	0.714	0.500	1.000	0.600	0.707	0.750	0.750
QDA	0.857	0.750	1.000	0.750	0.866	0.857	0.875
LR	0.714	0.500	1.000	0.600	0.707	0.750	0.750
AB	0.857	0.750	1.000	0.750	0.866	0.857	0.875
DT	0.571	0.750	0.333	0.500	0.500	0.400	0.542
BT	0.571	0.750	0.333	0.500	0.500	0.400	0.542
RF	0.714	0.500	1.000	0.600	0.707	0.750	0.750

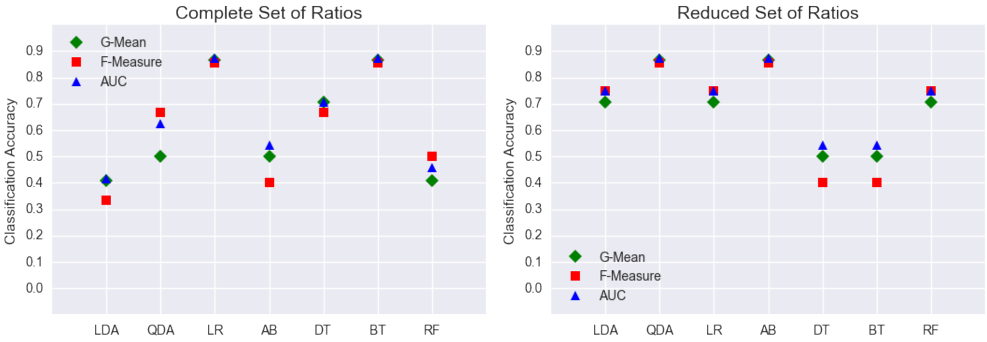


FIGURE 6.1: Models Classification Accuracy
Industry: Agriculture, Forestry and Fishing

Mining and Construction

Similar results can be seen for companies belonging to this industry (Table 6.3) when considering both scenarios, the complete set of ratios and the reduced one, which further supports the usefulness of variable selection.

Figure 6.2 expose the remarkable performance achieved by the random forests technique when all financial ratios are used, and by decision trees and QDA when the reduced set of variables is considered. In particular, the decision tree model correctly classifies 83.3% of non-fraudulent firms and 80% of fraudulent companies.

TABLE 6.3: Classification Accuracy Industry Specific
SIC 2: Mining and Construction

Sample size: 104							
	Accuracy	Specificity	Sensitivity	Precision	G-Mean	F-Measure	AUC
Complete set of financial ratios $K = 13$							
LDA	0.688	0.833	0.600	0.857	0.707	0.706	0.717
QDA	0.781	0.833	0.750	0.882	0.791	0.811	0.792
LR	0.625	0.750	0.550	0.786	0.642	0.647	0.650
AB	0.750	0.750	0.750	0.833	0.750	0.789	0.750
DT	0.781	0.833	0.750	0.882	0.791	0.811	0.792
BT	0.750	0.833	0.700	0.875	0.764	0.778	0.767
RF	0.875	0.917	0.850	0.944	0.883	0.895	0.883
Reduced set of financial ratios $k = 10$							
LDA	0.656	0.917	0.500	0.909	0.677	0.645	0.708
QDA	0.812	0.917	0.750	0.938	0.829	0.833	0.833
LR	0.688	0.917	0.550	0.917	0.710	0.687	0.733
AB	0.625	0.667	0.600	0.750	0.632	0.667	0.633
DT	0.812	0.833	0.800	0.889	0.816	0.842	0.817
BT	0.750	0.833	0.700	0.875	0.764	0.778	0.767
RF	0.781	1.000	0.650	1.000	0.806	0.788	0.825

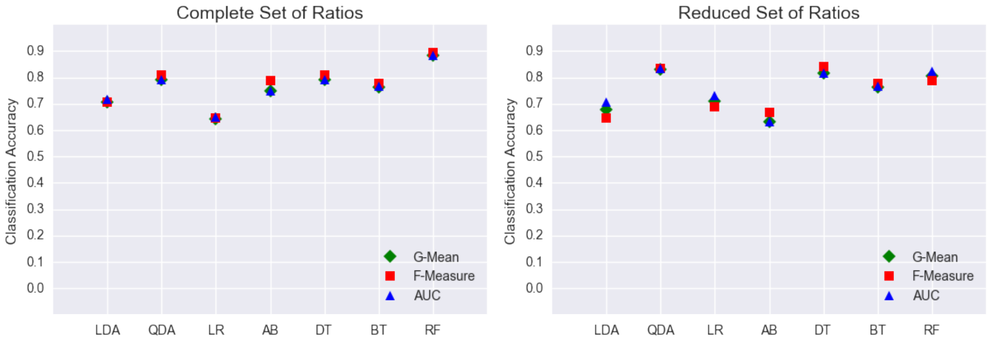


FIGURE 6.2: Models Classification Accuracy
Industry: Mining and Construction

Manufacturing

Moderate performance of the predictive models when dealing with manufacturing firms, as evidenced in Table 6.4. Similar results are obtained when considering the complete set of ratios and when considering the reduced one.

Figure 6.3 shows better results obtained by boosted trees when using all financial ratios, as well as AdaBoost when only the reduced set of ratios is considered. In this particular case, 55.7% of non-fraudulent cases and 60.9% of fraudulent cases are correctly classified.

TABLE 6.4: Classification Accuracy Industry Specific
SIC 3: Manufacturing

Sample size: 1,218	Accuracy	Specificity	Sensitivity	Precision	G-Mean	F-Measure	AUC
Complete set of financial ratios $K = 13$							
LDA	0.533	0.321	0.754	0.515	0.492	0.612	0.538
QDA	0.516	0.107	0.944	0.503	0.318	0.656	0.526
LR	0.536	0.337	0.743	0.518	0.500	0.610	0.540
AB	0.560	0.476	0.648	0.542	0.555	0.590	0.562
DT	0.566	0.829	0.291	0.619	0.491	0.395	0.560
BT	0.617	0.594	0.642	0.602	0.618	0.622	0.618
RF	0.549	0.401	0.704	0.529	0.531	0.604	0.552
Reduced set of financial ratios $k = 6$							
LDA	0.530	0.460	0.594	0.548	0.522	0.570	0.527
QDA	0.546	0.109	0.943	0.539	0.321	0.686	0.526
LR	0.530	0.425	0.625	0.545	0.516	0.583	0.525
AB	0.585	0.557	0.609	0.603	0.583	0.606	0.583
DT	0.555	0.259	0.823	0.551	0.461	0.660	0.541
BT	0.574	0.621	0.531	0.607	0.574	0.567	0.576
RF	0.503	0.460	0.542	0.525	0.499	0.533	0.501

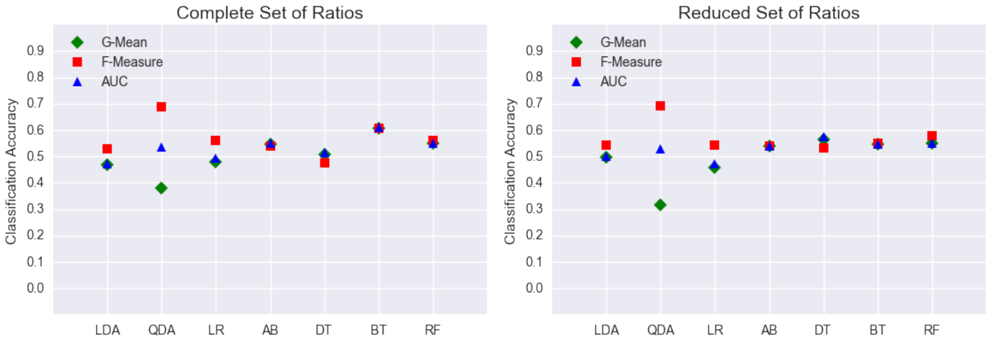


FIGURE 6.3: Models Classification Accuracy
Industry: Manufacturing

Transportation, Communications, Electric, Gas and Sanitary Service

Again, similar results can be seen for companies belonging to this industry when considering the complete set of ratios and the reduced set of ratios (Table 6.5), which reinforces the benefit of using only meaningful predictors. Furthermore, Figure 6.4 expose better performance achieved by boosted models, that is boosted trees, random forests and AdaBoost.

TABLE 6.5: Classification Accuracy Industry Specific
SIC 4: Transportation, Communications, Electric, Gas and Sanitary Service

Sample size: 212	Accuracy	Specificity	Sensitivity	Precision	G-Mean	F-Measure	AUC
Complete set of financial ratios $K = 13$							
LDA	0.625	0.688	0.562	0.643	0.622	0.600	0.625
QDA	0.625	0.938	0.312	0.833	0.541	0.455	0.625
LR	0.609	0.656	0.562	0.621	0.608	0.590	0.609
AB	0.672	0.750	0.594	0.704	0.667	0.644	0.672
DT	0.578	0.406	0.750	0.558	0.552	0.640	0.578
BT	0.656	0.688	0.625	0.667	0.656	0.645	0.656
RF	0.672	0.656	0.688	0.667	0.672	0.677	0.672
Reduced set of financial ratios $k = 5$							
LDA	0.562	0.625	0.500	0.571	0.559	0.533	0.562
QDA	0.562	0.969	0.156	0.833	0.389	0.263	0.562
LR	0.578	0.594	0.562	0.581	0.578	0.571	0.578
AB	0.609	0.719	0.500	0.640	0.599	0.561	0.609
DT	0.531	0.625	0.438	0.538	0.523	0.483	0.531
BT	0.672	0.719	0.625	0.690	0.670	0.656	0.672
RF	0.656	0.562	0.750	0.632	0.650	0.686	0.656

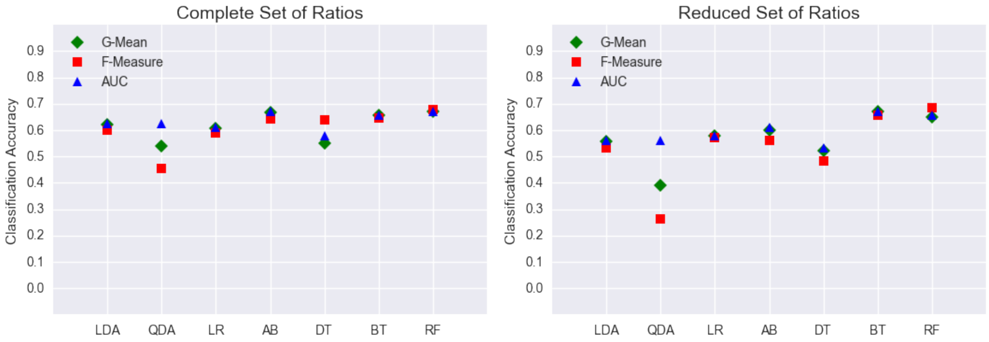


FIGURE 6.4: Models Classification Accuracy
Industry: Transportation, Communications, Electric, Gas and Sanitary Service

Wholesale and Retail Trade

In the case of trading firms, it can be observed from Table 6.6 that in most of the cases, the reduced set of ratios produce more accurate results compared to the complete set.

Figure 6.5 shows superior performance accomplished by boosted trees when detection both fraudulent and non-fraudulent companies. Decision trees in particular, achieved exceptional results when predicting fraudulent cases, but poor performance when dealing with non-fraud firms.

TABLE 6.6: Classification Accuracy Industry Specific
SIC 5: Wholesale and Retail Trade

Sample size: 338	Accuracy	Specificity	Sensitivity	Precision	G-Mean	F-Measure	AUC
Complete set of financial ratios $K = 13$							
LDA	0.490	0.562	0.426	0.523	0.489	0.469	0.494
QDA	0.520	0.271	0.741	0.533	0.448	0.620	0.506
LR	0.500	0.500	0.500	0.529	0.500	0.514	0.500
AB	0.618	0.604	0.630	0.642	0.617	0.636	0.617
DT	0.667	0.479	0.833	0.643	0.632	0.726	0.656
BT	0.637	0.646	0.630	0.667	0.638	0.648	0.638
RF	0.627	0.667	0.593	0.667	0.629	0.627	0.630
Reduced set of financial ratios $k = 3$							
LDA	0.559	0.521	0.593	0.582	0.556	0.587	0.557
QDA	0.500	0.042	0.907	0.516	0.194	0.658	0.475
LR	0.549	0.521	0.574	0.574	0.547	0.574	0.547
AB	0.608	0.542	0.667	0.621	0.601	0.643	0.604
DT	0.637	0.479	0.778	0.627	0.610	0.694	0.628
BT	0.745	0.771	0.722	0.780	0.746	0.750	0.747
RF	0.637	0.625	0.648	0.660	0.636	0.654	0.637

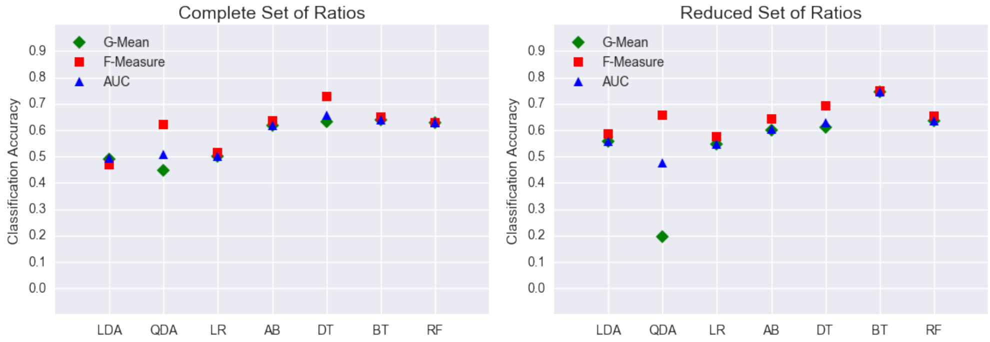


FIGURE 6.5: Models Classification Accuracy
Industry: Wholesale and Retail Trade

Finance, Insurance and Real Estate

Similar results can be seen from Table 6.7 when considering the complete set of ratios and the reduced one in the case of financial firms, again supporting the usefulness of using only meaningful predictors. Better performance accuracy is achieved by more advanced methodologies in both circumstances (Figure 6.6), particularly by the booted tree model, as it correctly classifies 68.2% of non-fraudulent cases and 63.8% of fraudulent cases.

TABLE 6.7: Classification Accuracy Industry Specific
SIC 6: Finance, Insurance and Real Estate

Sample size: 472	Accuracy	Specificity	Sensitivity	Precision	G-Mean	F-Measure	AUC
Complete set of financial ratios $K = 13$							
LDA	0.542	0.712	0.395	0.612	0.530	0.480	0.553
QDA	0.577	0.182	0.921	0.565	0.409	0.700	0.551
LR	0.535	0.606	0.474	0.581	0.536	0.522	0.540
AB	0.585	0.606	0.566	0.623	0.586	0.593	0.586
DT	0.592	0.424	0.737	0.596	0.559	0.659	0.581
BT	0.697	0.697	0.697	0.726	0.697	0.711	0.697
RF	0.606	0.697	0.526	0.667	0.606	0.588	0.612
Reduced set of financial ratios $k = 8$							
LDA	0.570	0.621	0.526	0.615	0.572	0.567	0.574
QDA	0.592	0.273	0.868	0.579	0.487	0.695	0.571
LR	0.570	0.591	0.553	0.609	0.571	0.579	0.572
AB	0.648	0.621	0.671	0.671	0.646	0.671	0.646
DT	0.627	0.561	0.684	0.642	0.619	0.662	0.622
BT	0.655	0.682	0.632	0.696	0.656	0.662	0.657
RF	0.627	0.652	0.605	0.667	0.628	0.634	0.628

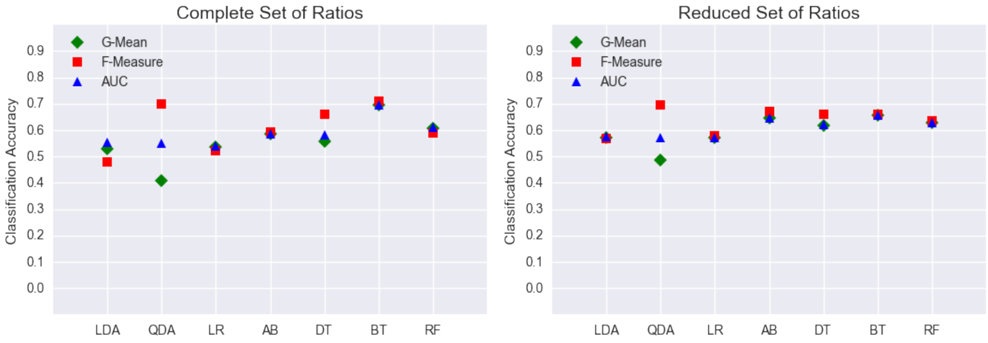


FIGURE 6.6: Models Classification Accuracy
Industry: Finance, Insurance and Real Estate

Services

Relatively good performance achieved by machine learning methods when detecting accounting fraud within the service industry, in particular when more advanced techniques are implemented, as evidenced in Figure 6.7.

One more time, similar results can be seen from Table 6.8 in both scenarios, that is, when considering the complete set of ratios and the reduced set, reinforcing the benefits of using only significant predictors.

TABLE 6.8: Classification Accuracy Industry Specific
SIC 7: Services

Sample size: 750	Accuracy	Specificity	Sensitivity	Precision	G-Mean	F-Measure	AUC
Complete set of financial ratios $K = 13$							
LDA	0.573	0.394	0.741	0.566	0.541	0.642	0.568
QDA	0.604	0.266	0.922	0.572	0.495	0.706	0.594
LR	0.547	0.413	0.672	0.549	0.527	0.605	0.543
AB	0.649	0.615	0.681	0.653	0.647	0.667	0.648
DT	0.618	0.578	0.655	0.623	0.615	0.639	0.617
BT	0.684	0.688	0.681	0.699	0.685	0.690	0.685
RF	0.676	0.606	0.741	0.667	0.670	0.702	0.673
Reduced set of financial ratios $k = 6$							
LDA	0.587	0.468	0.698	0.583	0.572	0.635	0.583
QDA	0.587	0.229	0.922	0.560	0.460	0.697	0.576
LR	0.587	0.495	0.672	0.586	0.577	0.627	0.584
AB	0.627	0.550	0.698	0.623	0.620	0.659	0.624
DT	0.631	0.615	0.647	0.641	0.630	0.644	0.631
BT	0.631	0.550	0.707	0.626	0.624	0.664	0.629
RF	0.618	0.477	0.750	0.604	0.598	0.669	0.614

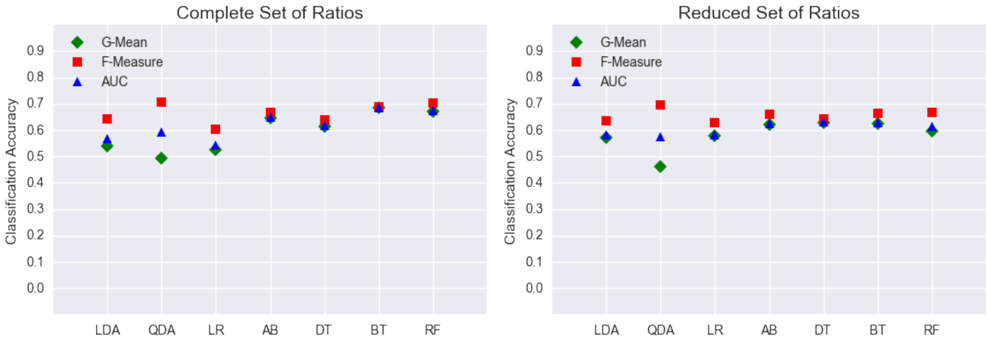


FIGURE 6.7: Models Classification Accuracy
Industry: Services

Public Administration

Exceptional results are obtained in the industry of public administration, as it is observed in Table 6.9 and Figure 6.8, in both scenarios, i.e.: when using the complete and the reduced set of financial ratios.

Particularly superior performance accomplished by random forests when the reduced set of explanatory variables are considered, as 90% of non-fraudulent cases are correctly classified, as well as 83.8% of fraudulent cases.

TABLE 6.9: Classification Accuracy Industry Specific
SIC 8: Public Administration

Sample size: 72	Accuracy	Specificity	Sensitivity	Precision	G-Mean	F-Measure	AUC
Complete set of financial ratios $K = 13$							
LDA	0.591	0.400	0.750	0.600	0.548	0.667	0.575
QDA	0.864	1.000	0.750	1.000	0.866	0.857	0.875
LR	0.727	0.600	0.833	0.714	0.707	0.769	0.717
AB	0.818	0.900	0.750	0.900	0.822	0.818	0.825
DT	0.818	0.900	0.750	0.900	0.822	0.818	0.825
BT	0.773	0.800	0.750	0.818	0.775	0.783	0.775
RF	0.818	0.900	0.750	0.900	0.822	0.818	0.825
Reduced set of financial ratios $k = 8$							
LDA	0.636	0.400	0.833	0.625	0.577	0.714	0.617
QDA	0.818	0.900	0.750	0.900	0.822	0.818	0.825
LR	0.727	0.600	0.833	0.714	0.707	0.769	0.717
AB	0.727	0.700	0.750	0.750	0.725	0.750	0.725
DT	0.773	0.700	0.833	0.769	0.764	0.800	0.767
BT	0.773	0.700	0.833	0.769	0.764	0.800	0.767
RF	0.864	0.900	0.833	0.909	0.866	0.870	0.867

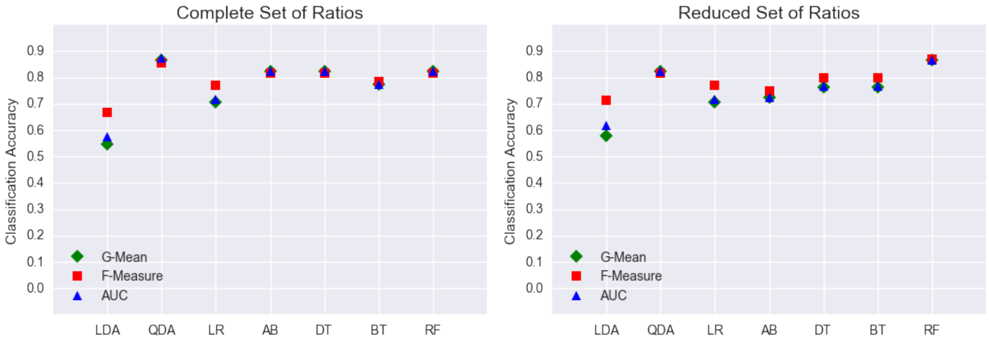


FIGURE 6.8: Models Classification Accuracy
Industry: Public Administration

6.4 Summary

Results obtained in Chapter 6, suggest that all machine learning methods proposed in this study provide superior predictive power compared to a naive strategy of classifying all firms as one class, either fraud or non-fraud. It is clear that explanatory variables used in developing a particular fraud detection model may differ from one industry to another, and therefore a domain-specific analysis is much more adequate when detecting accounting fraud offences.

It is worth mentioning as well, that in almost all industries, using the reduced set of financial ratios leads to similar results compared to the ones achieved using the complete set, in some cases even better performance is accomplished, which is very interesting since it supports the usefulness of employing less but significant information when detecting accounting fraud offences.

In the next chapter, decision rules obtained from decision tree models will be identified and further explained in order to responsibly assist the task of examination of financial statement reports. The proposed methodology is intended to support regulatory efforts to accurately detect, not only fraudulent corporations, but also financial accounts that are more likely to be manipulated.

Chapter 7

Financial Indicators of Accounting Fraud

In this Chapter, practical suggestions for effective examination of accounting information are given to further detect accounting fraud offences. Control strategies can be easily established taking into account different fraudulent tricks that are most commonly used when committing accounting fraud. What it is aimed in this case is to assist the oversight task related to accounting fraud offences using analytical results, to ultimately attempt to reduce the attractiveness of criminal opportunities within the corporate context.

Several statistical models have been implemented to better discriminate between fraudulent and non-fraudulent firms within different industry domains. Results from Chapter 6 support good performance of decision trees as a fraud-risk assessment tool. As such, rules obtained from these models can be described to expose industry-specific fraudulent behaviour, to be used later as warning signs of accounting fraud offences. Thereby, tree-based models will be interpreted in what follows to exhibit domain-specific financial indicators of accounting fraud.

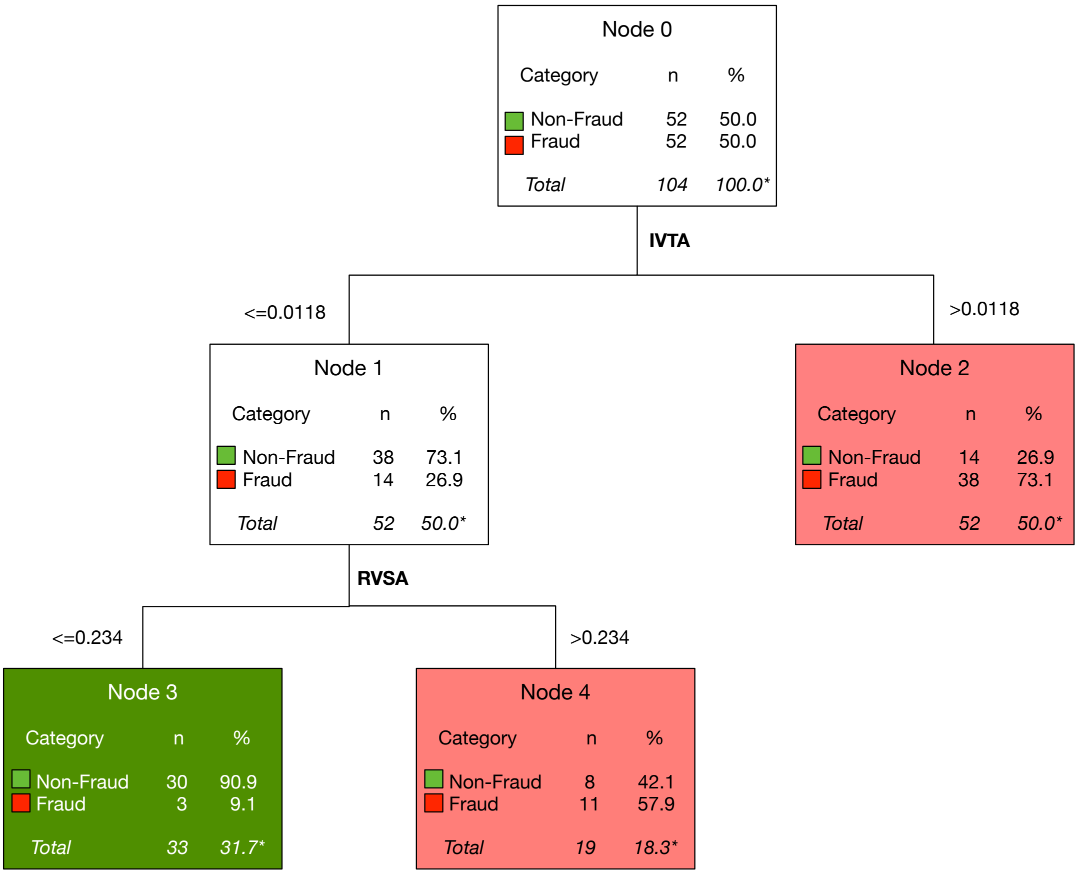
It is worth mentioning that no relevant patterns have been found within the Agriculture, Forestry and Fishing industry probably due to the small amount of available data, and therefore the inability of finding significant red flags in this domain.

7.1 Mining and Construction

As depicted in Figure 7.1, two main red-flags can be used to detect fraudulent companies belonging to the mining and construction industry that are specifically associated with the items of inventory and accounts receivable.

- The first indicator of accounting fraud is IVTA. The evidence suggests that it is more likely to be in presence of fraud when this ratio is bigger than 0.0118, which indicates that fraudulent firms tend to exaggerate inventory levels in this particular industry. Hence, fraud alarms should be activated when inventories represent more than 1.2% of total assets in mining and construction firms.
- The second indicator than can be used to expose falsified reports is RVSA. As such, when inventory levels compared to assets (IVTA) are within the non-fraudulent range (i.e.: lower than 0.0118), then auditors should check if RVSA levels are higher that 0.234. Therefore, the greater the probability of accounting fraud when figures of receivables represent more than 23.4% of total sales.

FIGURE 7.1: Decision Tree Visualisation
Industry: Mining and Construction



7.2 Manufacturing

Figure 7.2 illustrates a fairly more complex scheme commonly perpetrated by fraudulent firms within the manufacturing industry, as the resulting tree contains more branching.

Falsifying reports in this case, usually involves the manipulation of three financial items, that is, retained earnings, current assets and total liabilities. Moreover, decision tree results indicate that auditors should be more skeptical if RETA is higher that -0.292, CATA lower than 0.347 and TLTE higher than 1.132, since these three red-flags together are often seen when fraud is being committed in manufacturing firms.

In other words, high probability of accounting fraud will be present when:

- accounts receivables represent more than 39.2% of total assets;
- the proportion of current assets in relation to total assets is lower than 0.347; and
- total liabilities are 13.2% or higher than shareholders' equity.

7.3 Transportation, Communication, Electric, Gas and Sanitary Service

It can be seen from Figure 7.3 that the two most significant predictors of accounting fraud committed in this industry are IVSA and PYCOGS. As such, fraudulent reporting is more likely to be occurring as a result of misstatement of inventory levels and/or accounts payable figures.

As for the case of inventory manipulation, the warning sign is triggered when IVSA is lower or equal than zero. And as explained in Chapter 3.4.4, figures of inventory and total sales cannot be negative due to the lack of economic meaning. Then, the only possibility in this case is that inventories are zero. Consequently, auditors should be cautious when null inventories are part of financial statements as it may be a sign of accounting fraud.

On the other hand, if inventory levels are not null, then fraud alarm should be activated when accounts payable represent 28.2% of cost of good sold, as it may be indicating fraudulent activities.

FIGURE 7.2: Decision Tree Visualisation
Industry: Manufacturing

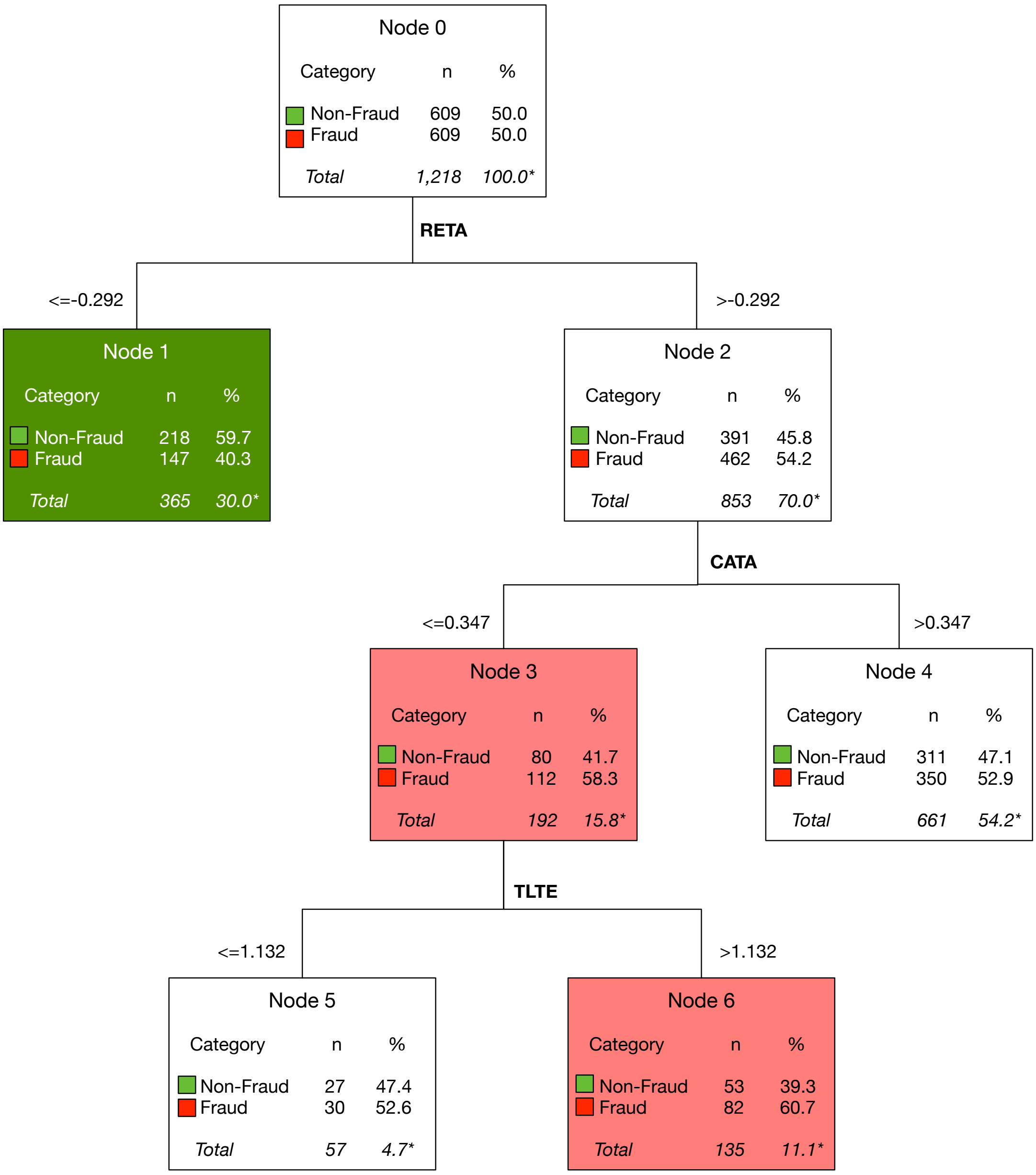
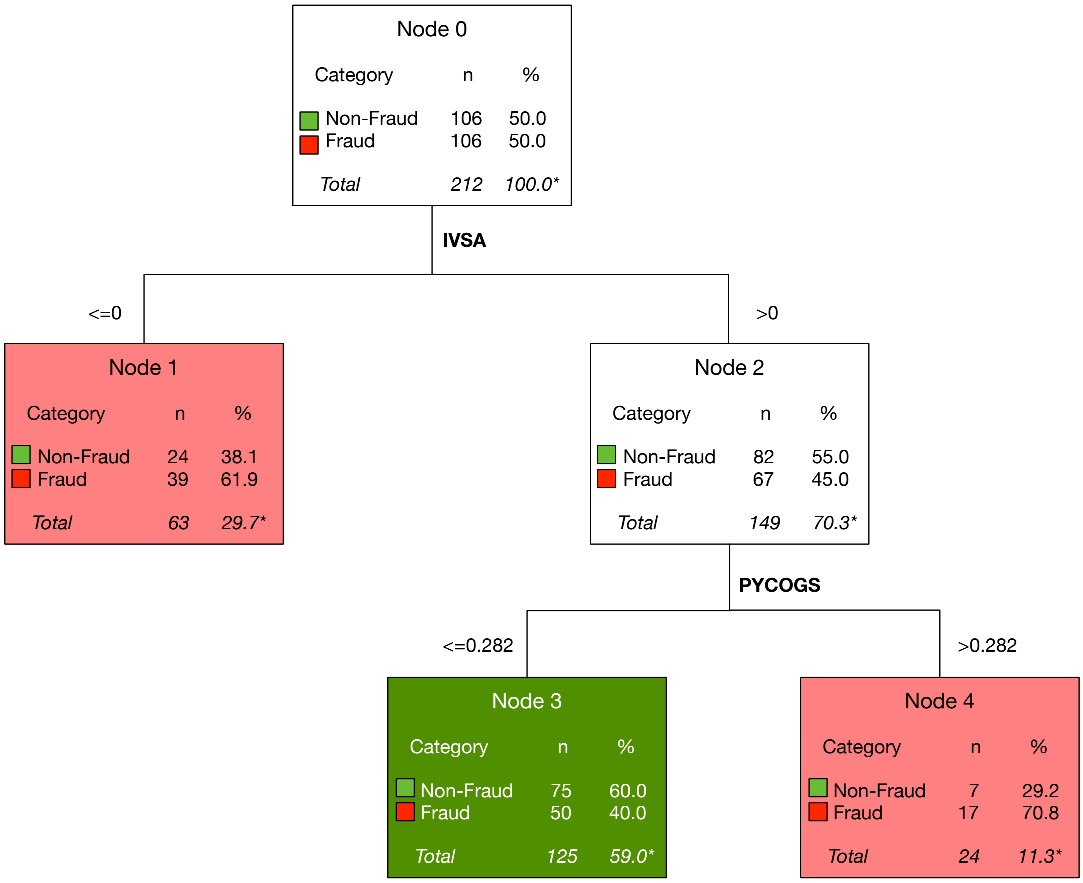


FIGURE 7.3: Decision Tree Visualisation
Industry: Transportation, Communication, Electric, Gas and Sanitary Service

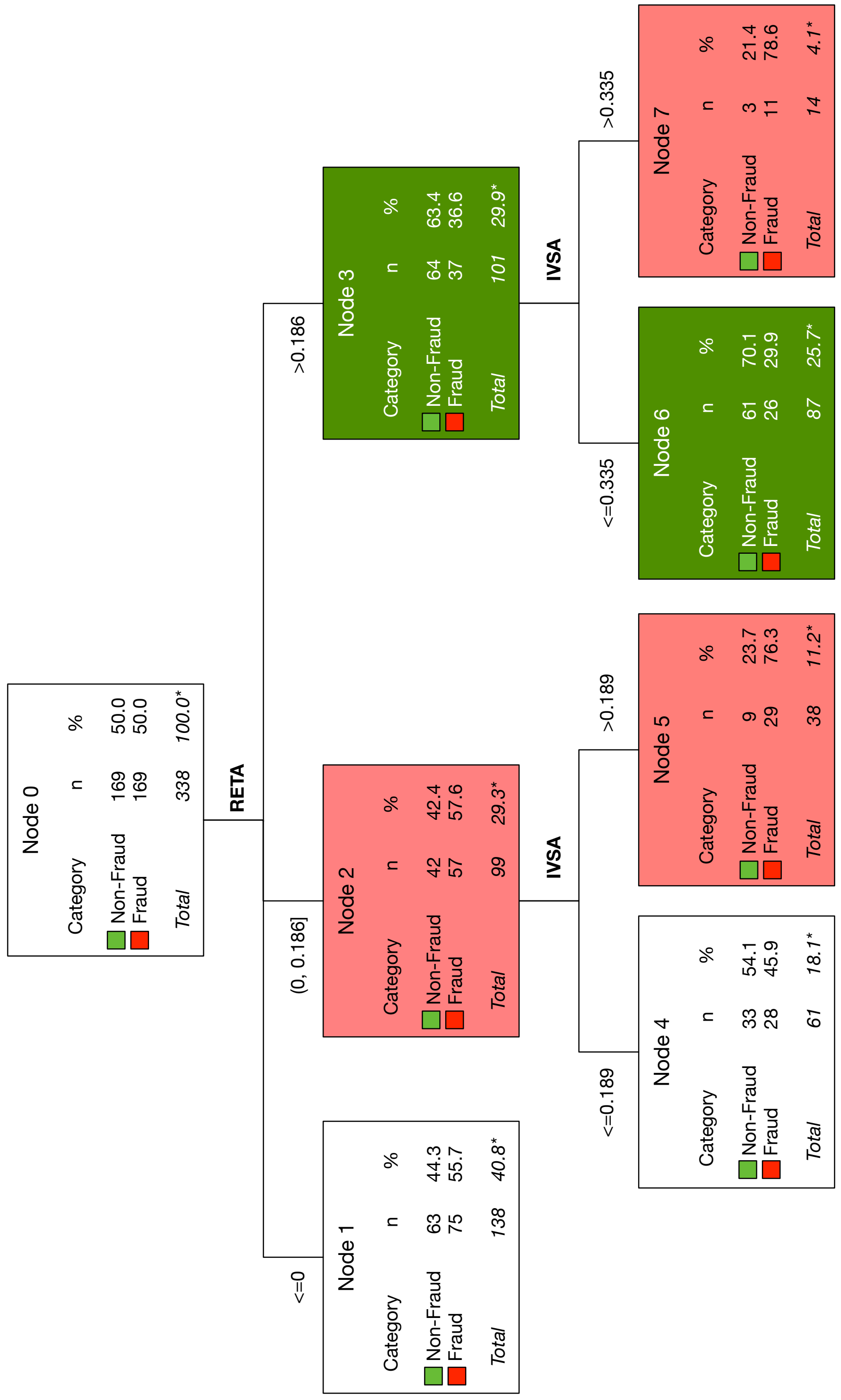


7.4 Wholesale Trade and Retail Trade

Results suggest that fraudulent trading companies manipulate mainly two financial items simultaneously, that is, retained earnings and inventories. Two clear patterns can be identified when accounting fraud is being committed, as shown in Figure 7.4.

- Financial ratio RETA between 0 and 0.186, as well as IVSA higher than 0.189. That is, moderate positive values of retained earnings and large values of inventory happening together represents a clear sign of falsified reports.
- Financial ratio RETA higher than 0.186 and at the same time, IVSA higher than 0.335 That is, exaggerated valuation of earnings compared to assets and inventory compared to sales are considered in this industry as irregular, hence more attention should be paid when facing this situation.

FIGURE 7.4: Decision Tree Visualisation
Industry: Wholesale Trade and Retail Trade



7.5 Finance, Insurance and Real Estate

It has been shown that the most significant predictors of accounting fraud committed in the finance industry are PYCOGS, LTDTA and TLTE. As seen in Figure 7.3, fraudulent reporting is more likely to be occurring as a result of manipulation of accounts payable and debt-specific figures.

On the one hand, if accounts payable are lower or equal to zero together with long-term debt higher than zero then more attention must be paid as it may be a sign of accounting fraud.

On the other hand, if accounts payable to cost of good sold are higher than 22.815 and, simultaneously, total liabilities are 19.05 times more than shareholders' equity, then warning alarm should be activated as irregular patterns are occurring that suggest fraudulent activities.

7.6 Services

As depicted in Figure 7.6, a fairly straightforward trick is usually performed by fraudulent companies in the industry of service, that is understating of sales figure together with the artificial exaggeration of inventory. More scrutiny should be made when total sales represent less than 25.6% of total assets, as well as when the proportion of inventory in terms of cost of good sold is higher than 0.032, as they may be indicating that accounting fraud is being conducted.

7.7 Public Administration

Accounting fraud in the industry of public administration is highly related to large values of inventory compared to sales, as it can be seen in Figure 7.7. Furthermore, special attention should be paid when evidencing inventories representing 6.3% or more of total sales, as this is a clear sign of manipulated financial reports.

FIGURE 7.5: Decision Tree Visualisation
Industry: Finance, Insurance and Real Estate

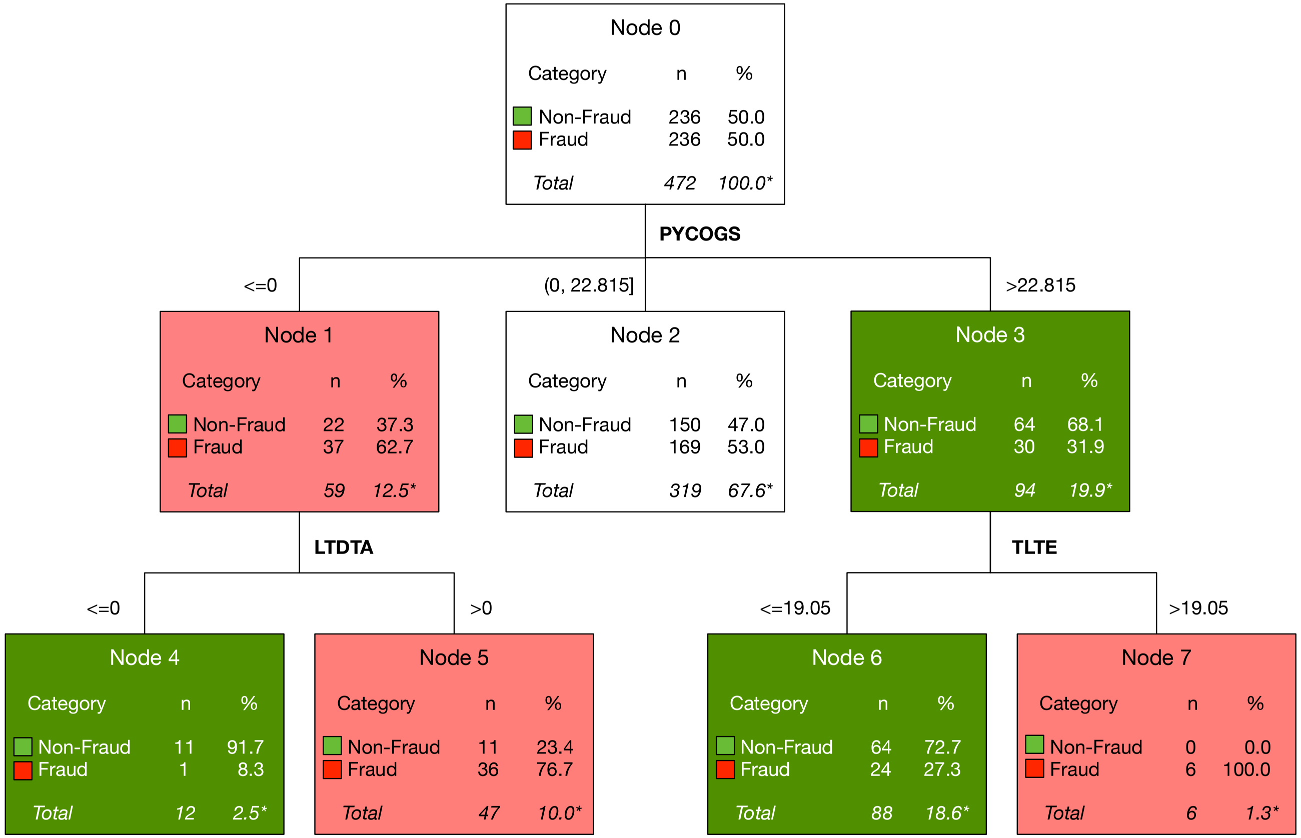


FIGURE 7.6: Decision Tree Visualisation
Industry: Services

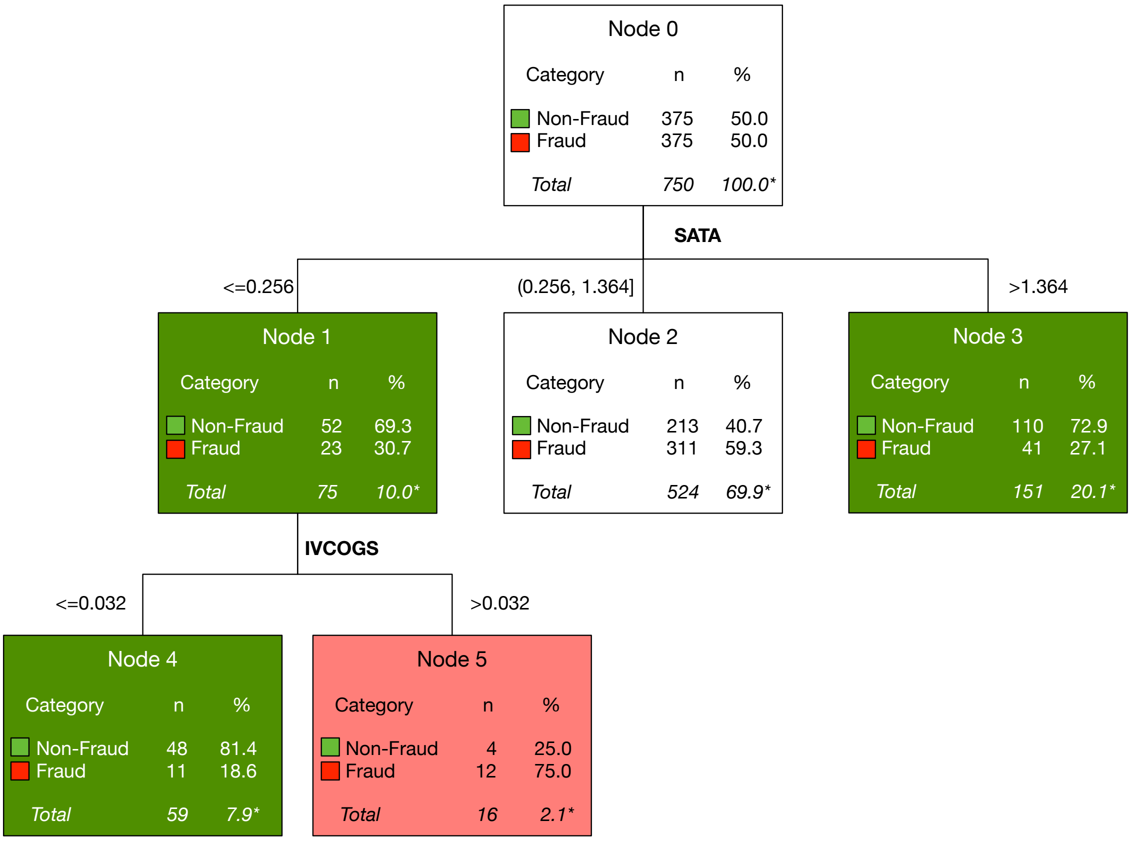
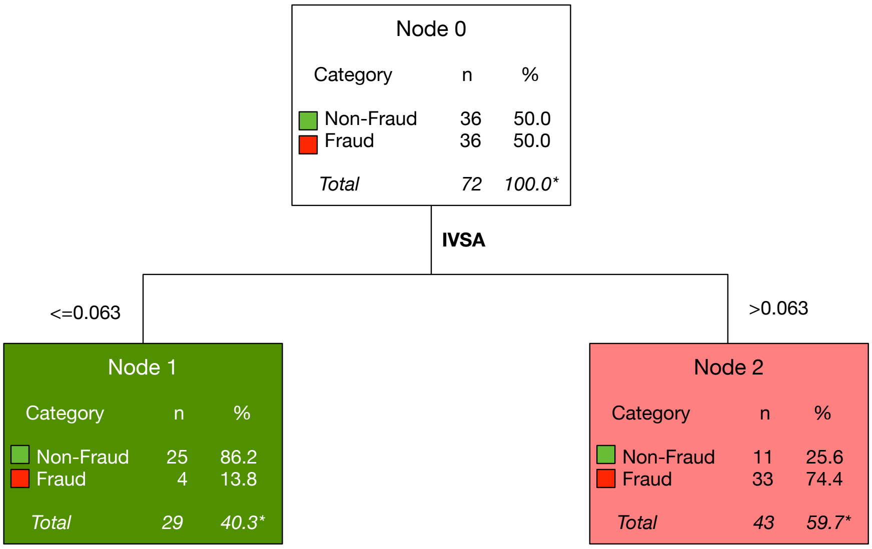


FIGURE 7.7: Decision Tree Visualisation
Industry: Public Administration



Chapter 8

Conclusions, Limitations and Future Work

8.1 Conclusions

This study aims to identify signs of accounting fraud occurrence to be used to, first, identify companies that are more likely to be manipulating financial statement reports, and second, assist the task of examination within the riskier firms by evaluating relevant financial red-flags, as to efficiently recognise irregular accounting malpractices.

To achieve this, a thorough forensic data analytic approach is proposed that includes all pertinent steps of a data-driven methodology. First, data collection and preparation is required to present pertinent information related to fraud offences and financial statements. Then, an in-depth financial ratio analysis is performed in order to analyse the collected data and to preserve only meaningful variables, selection that will be validated later using a more sophisticated technique that extends the well-known approach of complete subset regression. Finally, statistical modelling of fraudulent and non-fraudulent instances is performed by implementing several machine learning methods, followed by the extraction of distinctive fraud-risk indicators related to each economic sector.

This study contributes in the improvement of accounting fraud detection in several ways, including the collection of a comprehensive sample of fraud and non-fraud firms concerning all financial industries, an extensive analysis of financial information and significant differences between genuine and fraudulent reporting, selection of relevant predictors of accounting fraud, contingent analytical modelling for better differentiate between non-fraud and fraud cases, and identification of industry-specific indicators of falsified records.

There is a clear enhancement in the understanding of the fraud phenomenon by the implementation of financial ratio analysis, mainly due to the interesting exposure of distinctive characteristics of falsified reporting and the selection of meaningful ratios as predictors of accounting fraud, later validated using a combination of logistic regression models. Interestingly, using only significant explanatory variables leads to similar results obtained when no selection is performed. Furthermore, better performance is accomplished in some cases, which strongly supports the usefulness of employing less but significant information when detecting accounting fraud offences.

The results of the current research suggest there is a great potential in detecting falsified accounting records through statistical modelling and analysis of publicly available accounting information. It has been shown good performance of basic models used as benchmark - Discriminant Analysis and Logistic Regression-, and better performance of more advanced methods such as AdaBoost, Decision Trees, Boosted Trees and Random Forests. Results support the usefulness of machine learning models as they appropriately meet the criteria of accuracy, interpretability and cost-efficiency required for a successful detection methodology.

The proposed methodology can be easily used by public auditors and regulatory agencies in order to assess the likelihood of accounting fraud, and also to be adopted in combination with the experience and instinct of experts to lead to better examination of accounting reports.

In addition, the proposed methodological framework could be of assistance to many other interested parties, such as investors, creditors, financial and economic analysts, the stock exchange, law firms and to the banking system, amongst others.

8.2 Limitations

The collected sample of accounting fraud offences is considered to be only a fragment of the population of companies issuing fraudulent financial statement, as there is no guarantee that non-fraudulent firms are in fact legitimate observations until proven otherwise. Also, non-public companies are excluded from this study as the SEC only has jurisdiction over publicly traded companies.

It is worth noting that accounting fraud is very versatile, and as such, will always evolve in terms of deceptive tricks. Managers will adapt their fraudulent schemes in order to successfully commit fraud, hence results obtained in this study are exclusively consequence of the investigation of the collected data and different conclusions may be reach when considering an alternative source of information.

Lastly, models performances are not ideal in some scenarios mainly due to sample size, omitted variables and/or implemented techniques. Better accuracy would be likely achieved if different predictors were included in the analysis, such as stock information, corporate governance data, management quality, macro-economic information, competitors? financial data and more.

8.3 Future Work

It is strongly suggested the inclusion of other relevant information to help better understanding the accounting fraud phenomenon, which may consist of qualitative variables, including corporate governance information and inside trading data, as well as time-evolving features and industry-trending benchmarks. It would not be surprising to discover interesting temporal patterns of stock prices or asset returns when dealing with fraudulent corporations, or find an extraordinary economic performance of dishonest companies compared to the industry average.

Further work can be done for classification threshold selection. When modelling the accounting fraud phenomenon, it was mentioned that a specific classification threshold was considered to determine fraud and non-fraud categories in several machine learning techniques. Evaluation of different thresholds would be of much interest as it may improve classification accuracy in a cost-sensitive environment such as the one at issue.

In addition, different methodologies are suggested to tackle the imbalance class challenge. The method adopted in the present study was based on random under-sampling, but other techniques may improve this part of the process, such as random over-sampling, bootstrap models, cost modifying methods and algorithm-level approaches, to name a few.

More advanced techniques are also recommended specifically when dealing with accounting fraud affairs, including missing value treatment, sample selection and imbalanced database issues, outlier detection and treatment, and variable selection. Additionally, it would be very interesting to implement alternative and more advanced machine learning methods, such as support vector machines, neural networks and Bayesian models, as they may be helpful to correctly identify fraudulent firms.

In addition, it is suggested to replicate the proposed methodology in specific economic domains, such as the pharmaceutical industry, health care industry and financial industry, amongst others. The more specialised the industry, the more interesting patterns are likely to be found.

Finally, more extensive analyses can be performed to tackle the broader topic of white-collar and corporate crimes, such as social network techniques that can be used to uncover sophisticated fraudulent networks between business affiliates and subsidiaries, as well as connection with other companies or even further associations with political bodies and governmental servants.

Bibliography

Altman, Edward (1968). “Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankrupcy”. In: *The Journal of Finance*.

Baesens, Bart, Veronique Van Vlasselaer, and Wouter Verbeke (2015). *Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques: A Guide to Data Science for Fraud Detection*. John Wiley and Sons, Inc.

Beasley, Mark (1996). “An Empirical Analysis of the Relation between the Board of Director Composition and Financial Statement Fraud”. In: *The Accounting Review*.

Bell, Timothy and Joseph Carcello (2000). “A Decision Aid for Assessing the Likelihood of Fraudulent Financial Reporting”. In: *Auditing: A Journal of Practice & Theory*.

Benson, Michael L. and Sally S. Simpson (2014). *Understanding White-Collar Crime: An Opportunity Perspective*. Criminology and Justice Studies. Taylor & Francis.

Bishop, Chistopher M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Bolton, Richard and David Hand (2002). “Statistical Fraud Detection: A Review”. In: *Statistical Science*.

Bose, Indranil, Selwyn Piramuthu, and Michael J. Shaw (2011). “Quantitative Methods for Detection of Financial Fraud”. In: *Decision Support Systems*.

Braithwaite, John (1984). *Corporate Crime in the Pharmaceutical Industry*. Routledge & Kegan Paul.

Cerullo, Michael and Virginia Cerullo (1999). “Using Neural Networks to Predict Financial Reporting Fraud”. In: *Computer Fraud and Security*.

Chawla, Nitesh V., Nathalie Japkowicz, and Aleksander Kotcz (2004). “Editorial: Special Issue on Learning from Imbalanced Data Sets”. In: *ACM SIGKDD Explorations Newsletter*.

Choi, Jae and Brian Green (1997). “Assessing the Risk of Management Fraud Through Neural Network Technology”. In: *Auditing*.

Crask, Melvin and William Perreault (1977). “Validation of Discriminant Analysis in Marketing Research”. In: *Journal of Marketing Research*.

Dalnial, Hawariah et al. (2014). “Detecting Fraudulent Financial Reporting through Financial Statement Analysis”. In: *Journal of Advanced Management*.

Elliot, Graham, Antonio Gargano, and Allan Timmerman (2013). “Complete Subset Regressions”. In: *Journal of Econometrics*.

– (2015). “Complete Subset Regressions with Large-Dimensional Sets of Predictors”. In: *Journal of Economic Dynamics and Control*.

Fanning, Kurt and Kenneth Cogger (1998). “Neural Network Detection of Management Fraud Using Published Financial Data”. In: *International Journal of Intelligent Systems in Accounting, Finance & Management*.

Feroz, Ehsan Habib et al. (2000). “The Efficacy of Red Flags in Predicting the SEC’s Targets: An Artificial Neural Networks Approach”. In: *International Journal of Intelligent Systems in Accounting, Finance & Management*.

Gupta, Rajan and Nasib Singh Gill (2012). “Prevention and Detection of Financial Statement Fraud - An Implementation of Data Mining Framework”. In: *International Journal of Advanced Computer Science and Applications*.

Hansen, James V. et al. (1996). “A Generalized Quanlitative-Response Model and the Analysis of Management Fraud”. In: *Management Science*.

Hollander, Myles, Douglas A. Wolfe, and Eric Chicken (2013). *Nonparametric Statistical Methods: Third Edition*. Wiley Series in Probability and Statistics. Wiley.

James, Gareth et al. (2013). *An Introduction to Statistical Learning*. Springer-Verlag New York.

Kaminski, Kathleen A., T. Sterling Wetzel, and Liming Guan (2004). “Can Financial Ratios Detect Fraudulent Financial Reporting?” In: *Managerial Auditing Journal*.

Kendall, Maurice G. (1955). *Rank Correlation Methods*. Hafner Publishing Co.

Kirkos, Efstathios, Charalambos Spathis, and Yannis Manolopoulos (2007). “Data Mining Techniques for the Detection of Fraudulent Financial Statements”. In: *Expert Systems with Applications*.

Kotsiantis, Sotiris et al. (2006). “Forecasting Fraudulent Financial Statements Using Data Mining”. In: *International Journal of Computational Intelligence*.

Kwon, Taek Mu and Ehsan Feroz (1996). “A Multilayered Perceptron Approach to Prediction of the SEC’s Investigation Targets”. In: *IEEE Transactions on Neural Networks*.

Lee, Thomas A., Robert W. Ingram, and Thomas P. Howard (1999). “The Difference Between Earnings and Operating Cash Flow as an Indicator of Financial Reporting Fraud”. In: *Contemporary Accounting Research*.

Lenard, Mary Jane, Ann Watkins, and Pervaiz Alam (2007). "Effective Use of Integrated Decision Making: An Advanced Technology Model for Evaluating Fraud in Service-Based Computer and Technology Firms". In: *Journal of Emerging Technologies in Accounting*.

Lin, Jerry, Mark Hwang, and Jack Becker (2003). "A Fuzzy Neural Network for Assessing the Risk of Fraudulent Financial Reporting". In: *Managerial Auditing Journal*.

Mokhiber, Russell (2007). "Twenty Things You Should Know About Corporate Crime". In: Taming the Giant Corporation Conference, Washington, D.C.

Mokhiber, Russell and Robert Weissman (2005). *On The Rampage: Corporate Power and the Destruction of Democracy*. Corporate Focus Series. Common Courage Press.

Ngai, Eric W.T. et al. (2011). "The Application of Data Mining Techniques in Financial Fraud Detection: A Classification Framework and an Academic Review of Literature". In: *Decision Support Systems*.

Ou, Jane A. and Stephen H. Penman (1989). "Financial Statement Analysis and the Prediction of Stock Returns". In: *Journal of Accounting and Economics*.

Pai, Ping Feng, Ming Fu Hsu, and Ming Chieh Wang (2011). "A Support Vector Machine-Based Model for Detecting Top Management Fraud". In: *Knowledge-Based Systems*.

Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research*.

Persons, Obeua S. (1995). "Using Financial Statement Data to Identify Factors Associated with Fraudulent Financial Reporting". In: *Journal of Applied Business Research*.

Ravenda, Diego, Josep M. Argilés-Bosch, and Maika M. Valencia-Silva (2015). "Detection Model of Legally Registered Mafia Firms in Italy". In: *European Management Review*.

Ravisankar, Pediredla et al. (2011). "Detection of Financial Statement Fraud and Feature Selection Using Data Mining Techniques". In: *Decision Support Systems*.

Savona, Ernesto and Giulia Berlusconi (2015). "Organized Crime Infiltration of Legitimate Businesses in Europe: A Pilot Project in Five European Countries". In: *Final Report of Project ARIEL – Assessing the Risk of the Infiltration of Organized Crime in EU MSs Legitimate Economies: a Pilot Project in 5 EU Countries*.

Sheskin, David J. (2003). *Handbook of Parametric and Nonparametric Statistical Procedures: Third Edition*. CRC Press.

Shilit, Howard M. and Jeremy Perler (2010). *Financial Shenanigans*. Mc Graw Hill.

Simpson, Sally S. (2002). *Corporate Crime, Law, and Social Control*. Cambridge Studies in Criminology. Cambridge University Press.

Simpson, Sally S. and David Weisburd, eds. (2009). *The Criminology of White-Collar Crime*. Springer-Verlag New York.

Song, Xin Ping et al. (2014). "Application of Machine Learning Methods to Risk Assessment of Financial Statement Fraud: Evidence from China". In: *Journal of Forecasting*.

Spathis, Charalambos (2002). "Detecting False Financial Statements Using Published Data: Some Evidence From Greece". In: *Managerial Auditing Journal*.

Spathis, Charalambos, Michael Doumpos, and Constantine Zopounidis (2002). "Detecting Falsified Financial Statements: A Comparative Study Using Multicriteria Analysis and Multivariate Statistical Techniques". In: *The European Accounting Review*.

Summers, Scott and John Sweeney (1998). "Fraudulently Misstated Financial Statements and Insider Trading: An Empirical Analysis". In: *The Accounting Review*.

Sutherland, Edwin H. (1949). *White Collar Crime*. Dryden Press.

Swartz, Mimi (2003). *Power Failure: The Inside Story of the Collapse of Enron*. Doubleday.

Tu, Jack (1996). "Advantages and Disadvantages of Using Artificial Neural Networks Versus Logistic Regression for Predicting Medical Outcomes". In: *Journal of Clinical Epidemiology*.

Van Vlasselaer, Veronique et al. (2015). "GOTCHA! Network-based Fraud Detection for Social Security Fraud". In: *Management Science*.

Wang, Shiguo (2010). "A Comprehensive Survey of Data Mining-based Accounting-Fraud Detection Research". In: *International Conference on Intelligent Computation Technology and Automation*.

Yarnold, Paul, Robert Soltysik, and Gary Martin (1994). "Heart Rate Variability and Susceptibility for Sudden Cardiac Death: An Example of Multivariable Optimal Discriminant Analysis". In: *Statistics in Medicine*.

Yarnold, Paul et al. (1995). "Application of Multivariable Optimal Discriminant Analysis in General Internal Medicine". In: *Journal of General Internal Medicine*.