‘BIG DATA ANALYTICS’ FOR CONSTRUCTION FIRMS INSOLVENCY PREDICTION MODELS

By

HAFIZ A. ALAKA

13044174

A thesis submitted in partial fulfilment of the requirements of the University of the West of England, Bristol for the degree of Doctor of Philosophy

Faculty of Business and Law, The University of the West of England, Bristol

January 2017

ABSTRACT

In a pioneering effort, the research was the first to develop a construction firms insolvency prediction model (CF-IPM) with: Big Data Analytics (BDA); combined qualitative and quantitative variables; advanced artificial intelligence tools such as Random Forest and Bart Machine; and data of all sizes of construction firms (CF), ensuring wide applicability

The pragmatism paradigm was employed to allow the use of mixed methods. Top management team (TMT) of existing and failed CFs were interviewed. This included large, medium, small and micro (LMSM) CFs. The interview result was used to create a questionnaire with over hundred qualitative variables. A total of 531 usable questionnaires were returned, and oversampled to a total questionnaire sample of 1052 LMSM CFs. The financial data of the original and matched sample firms were downloaded. Using Cronbach's alpha and factor analysis, qualitative variables were reduced to 13. Eleven financial ratios commonly reported by the sample LMSM CFs were identified as quantitative variables. Using BDA, implemented through Amazon Web Services Elastic Compute Cloud, eleven variable selection methods were used to select the final seven qualitative and seven quantitative variables which were used to develop 13 BDA-CF-IPMs.

A key finding was that the Decision Tree BDA-CF-IPM was the best model because it had high accuracy and was transparent enough to show where a potentially failing CF was deficient. Also, results showed that the normally high performing artificial neural network and support vector machine AI tools were not good at handling a combination of quantitative and qualitative variables. On the variables part, a key discovery was that while high immigration levels favour large CFs, it is a major challenge to medium, small and especially, micro (MSM) CFs.

A key achievement and contribution was the successful implementation of BDA to develop CF-IPMs, eliminating the problem of long development days due to high computation intensity. Another achievement was the development of CF-IPMs with extreme accuracy levels of over 99% using contemporary AI tools. Also, the adopted methodology helped to contribute potential qualitative variables for interested future CF-IPM studies. Finally, the developed model was the first CF-IPM applicable to all sizes of CFs, including the MSM CFs which make up over 90% of the construction industry.

DEDICATION

This project work, as usual, is firstly dedicated to Almighty ALLAH the LORD of the world, the Beneficent and the Merciful. I thank you for leading me to complete this project work successfully. Further dedication goes to my late mum for her unrelenting continuous support of all types when she was alive. May ALLAH (SWT) grant her Al-Jannah Firdaus. Words cannot express my feelings. Next in line is my dad. I thank him for his continuous gigantic financial and moral support. May ALLAH (SWT) continue to guide, guard, bless and preserve him. May He give you long life and prosperity and grant him Al-Jannah Firdaus in the hereafter. Ameen.

Next in line is my golden icons in persons of my wife Hidaya Olajumoke Alaka, and my children, Yusayrah and Mahfouz Alaka. Your continuous moral, lovely and unending support cannot be overrated. I love you and will always love you. May ALLAH continue to guide and guard us, give us long life and prosperity and ultimately grant us all Al-Jannah Firdaus in the hereafter. Ameen

Left to follow are my brothers in persons of Lukman, Waheed, Mustapha, Sulaiman, Abdullah and Waliu, and my sisters in persons of Bilkis, Maryam and Lateefah. I cannot imagine life without your continuous moral and financial support. I cannot thank you all enough. I love you all. Next on the list are my nieces Maryam, Zainab, Aisha, Rahma, Raqia and Fatimah, and my nephews Azeez, Jamil (don), Hameed, Awwal, Abubakar and Muqaffi. I love you all with your massive troublesome moves. May ALLAH continue to guide and guard our family, give us long life and prosperity and ultimately grant us Al-Jannah Firdaus in the hereafter. Amin

Lastly are my late step brothers Yussuf and Usman and my late grandmother Alhaja Safurat Owodunni. May ALLAH (SWT) grant you and all Muslims Al-Jannah Firdaus. Ameen.

ACKNOWLEDGEMENT

First and foremost, I acknowledge the grace of Almighty ALLAH that He has bestowed on me. Without such grace, this work could not have been completed. I thank Him, I praise Him, I adore Him.

My unending appreciation then goes to my supervisor in person of Professor Lukumon Oyedele. He is God sent! He has been a wonderful guide and director on this project and beyond. He has been an inspiration in every aspect of my post-graduate academic life. I cannot thank him enough for his efforts. If I were given a chance, I would choose him as my supervisor time and again. I also acknowledge the support and guidance received from my second supervisor, Dr Vikas Kumar. His contributions were immeasurable.

I acknowledge and greatly appreciate the moral and financial support given to me by my family including my immediate nuclear family, my parents and my siblings

I acknowledge the great contributions of my colleagues in persons of Owolabi Hakeem, Ajayi Saheed, Muhammed Bilal and Akinade Oluwagbenga, Together, we formed a formidable team of early career academics with various great achievements under the guidance of Professor Lukumon Oyedele. You will all be sorely missed (including Professor Lukumon).

Lastly, I will like to acknowledge the efforts of Dr Svetlana Cicimil as the Director of Doctoral Research. She was wonderful. This acknowledgement will not be complete without mentioning the staff at the graduate school office, especially Dr Helen Frisby, Samantha Watts and Paul spencer. Their guidance, timely reminders, skills development programmes, among others, were instrumental to a smooth programme.

Overall I thank Allah whom all praise belongs to, for seeing me through. Nothing at all could have been achieved without His consent. ‘Alhamdulillahi Robbil Alamin’.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

1.0 INTRODUCTION

1.1 Background

The construction industry (CI) is a vital part of most countries' economy (Zhao *et al.*, 2012). On the global scale, the CI had a staggering worth of US$7.4 trillion in 2010, has a projection of US$10.3 trillion in 2020 (Department for Business Innovation and Skills, 2013a) and $15.5 trillion by 2030 (Global Construction Perspectives and Oxford Economics, 2015).

According to the (Department for Business Innovation and Skills, 2013a), the CI in 2013 was responsible for about £90 billion or 6.7 percent of the United Kingdom (UK) economy and covered more than 280,000 businesses, providing more than two million jobs. According to Rhodes (2015) in a House of Commons Library research paper, the CI in 2014 contributed £103 billion in economic output, representing 6.5% of the total; it also provided 2.1 million jobs or 6.2% of the UK total in 2015. The continuous mass failure witnessed in the construction industry thus cause real economic troubles, showing the need for improvement on insolvency prediction models. Such models are needed by construction firms for self-assessment, and by clients and financial institutions to ensure contracts and loans respectively are given to healthy firms.

The record of firm failure in the UK CI is alarming. To mention a few, AD Utting Construction Limited, Duart Construction, TRS Services, Team Project Limited, Kitpac Building, Sunnydale Civil Engineering, Colin Amos Builders and John Kotes Construction and Site Services Limited, are just some of the companies that became insolvent in UK in December 2010 alone (The Construction Index, 2011). In 2012, a company insolvency ranking by sector showed that the UK construction sector ranked the third highest (at 14.4 percent) among other business sectors in the UK. Overall, the industry has consistently featured among the top three in UK insolvency ranking by sector over the years including in the latest reports of 2016, where it ranks as number one (The Insolvency service, 2016).

Researchers have attributed construction firms' relative high failure rate to risks such as fluctuation in demand, policy changes affecting the economy, fluctuating cost of materials, high rate of litigation, safety issues, cash flow problems, among others (Mason and Harris 1979; Ng et al. 2011; Chen 2012). Enshassi *et al.* (2006) believe the main problem is the ease with which companies get into the industry, causing an influx which results in fierce competition thereby leading to a soaring rate of business failure. The Surety Information Office (2012), in their review, concluded that the main causes of construction firms' failure are their low-performance level, illusory growth, account problems, character and management issues. With these identified risks and the potential complexity of managing them, the high volume of insolvency in the CI almost excuses itself. Nonetheless, high rate of insolvency is not something any industry can afford to live with. These identified risks (most are non-financial) also clearly indicate, as supported by many studies, that financial indicators alone cannot be used to identify a potential insolvency early enough. The truth is it is company/managerial activities, performance and characteristics that result in the financial situation of a firm (Abidali and Harris 1995) (see chapter three for more on indicators/variables).

The fact that the record of failed firms in the construction industry is alarming and somehow proportionate to the risks involved makes it look like nothing is being done to tackle the rate of insolvency, yet some insolvency prediction models have been developed over time to help avoid insolvency. *There are however two major questionable areas of insolvency prediction models built for construction firms: 'data' and 'tool'.*

On data issue, the problem is that studies that build insolvency prediction models for the CI rely mainly on financial statements of the sample firms (e.g. Mason and Harris 1979; Langford *et al.* 1993; Chen 2012; Bal *et al.* 2013; Horta and Camanho 2013 among others). This step is in blind followership of pioneer insolvency prediction models (IPM) studies (i.e. Beaver 1966; Altman 1968) that had their own unique objectives that perfectly allowed exclusive use of financial ratios. "*A financial ratio is a quotient of two numbers, where both numbers consist of financial statement items*" (Beaver 1966; pp. 71-72). This method is inadequate as it fails to properly take into account small and micro enterprise (SME) construction firms which represent over 97% of the UK construction industry (Department for Business Innovation and Skills, 2015). So how is this method inadequate? The answers are: (i) The method excludes firms with incomplete accounting data, a major feature of SMEs, from their model. (ii) Where SMEs have complete accounting data, they are usually misrepresenting statements because SMEs frequently outsource financial statement production with the main aim of meeting the legal requirement of annual production

(Balcaen and Ooghe 2006). (iii) Some SMEs simply do not produce statements at all. These answers/facts, coupled with long ago established the importance of non-financial variables (Argenti, 1980; Zavgren, 1985), clearly show the need for a robust model that combines financial (quantitative) and non-financial (especially qualitative) variables to be built for construction firms.

On tools issue, following recommendations in Beaver's Beaver (1966) univariate prediction model, Altman (1968) and Ohlson (1980) used multi-discriminant analysis (MDA) and logit analysis (LA) tools respectively to develop multivariate prediction models. These models that were more accurate than Beaver's, widely accepted and improved over time. These tools were subsequently well applied to firms' failure research in the CI (Mason and Harris 1979; Langford *et al.* 1993; Abidali and Harris 1995; Ng *et al.* 2011; Bal *et al.* 2013). The various problems of these statistical tools, however, led to the rise and wider acceptance of artificial intelligence (AI) tools as their replacement (see section 5.2.1 for more on tools). Nonetheless, very limited studies have used AI tools to develop insolvency prediction models for the CI (e.g. Horta and Camanho 2013; Chen 2012 among others); in fact, *no study has used AI tools to develop a model for the UK CI*.

Further, despite clear evidence that large data improves reliability, only very few IPM studies (e.g. Altman *et al.* 1994; Du Jardin 2010) have been able to use a good size of data with AI tools. Although Altman *et al.* 1994 used a relatively large sample of 1000 firms with artificial neural network (ANN), the parameters of the ANN were not tuned, which means the model achieved is not the optimum achievable with ANN. Du Jardin (2010) used a smaller data set of 500 firms but tuned the (topology, learning rate, momentum term, weight decay) parameters of ANN, leading to a higher computational intensity and probably a much better model. As a result, "it took roughly five days to compute all network parameters with 30 PCs running Windows, and an additional day to calculate and check the final results" (Du Jardin 2010; p.2052). Contemporary technology such as Big Data Analytics can help to avoid such a tedious computation duration without sacrificing the necessary parameters tuning.

Overall, the shortcomings of the developed models include: (i) The over-reliance on financial indicators i.e. financial ratios for building the models thereby maligning the more important SME construction firms. (ii) The refusal to adopt the most contemporary technology like well-tuned high performing AI tools and Big Data Analytics that can be used to build more sophisticated models. These shortcomings and the continuing high construction business failures as earlier identified only emphasise the need for better robust insolvency prediction models.

1.2 Concept of business insolvency

In Wales and England, the word bankruptcy applies only to individuals and is governed by the Insolvency Rules 1986 (as amended) and Part IX of the Insolvency Act 1986 (as amended). Insolvency, which is more of the bankruptcy term for limited companies in the UK, is regulated by United Kingdom insolvency law which can easily call for compulsory liquidation (Gov.uk, 2014). It is supported mainly by the Insolvency Act 1986 and the Insolvency Rules 1986; both include numerous subs which can only be valid by a court sentence/order, making it very hard to have a single bankruptcy/insolvency definition.

While Beaver's (1966) seminal work gave a broad definition of insolvency, Scott (1981) noted that many subsequent studies limited their definitions to bankruptcy. Watson and Everett (1993) simple definition says that insolvency of business happens when one of four circumstances occur: (i) ending the business for any reasons; (ii) termination of trading and losses of credit; (iii) selling of business to avoid more losses; (iv) and not successfully starting the business

In the research, insolvency will be defined as any firm that has gone through insolvency through receivership or liquidation under the United Kingdom insolvency law. These measures are straightforward, and this sort of firms are easily identifiable from the databases where financial data will be extracted for analysis. Although there are other types of distresses which may include informal support from the government, renegotiating loan terms, merging with a more stable firm, acute downsizing, among others, these are quite difficult to identify and are thus not considered in this work. The terms failure, bankruptcy and insolvency are used interchangeably throughout this thesis write up.

1.3 Big data and its concept

John Mashey was the first person to use the set of words ‘Big Data’ together when he did a presentation on Silicon Graphics (SGI) slide titled “Big Data and the Next Wave of InfraStress” (Diebold, 2012a). The relativity of the word ‘big’ makes the definition of ‘big data’ complex but a generally accepted fact is that for data to qualify as big data, it must have either or all of three characteristics namely: velocity, volume and variety (Zikopoulos and Eaton 2011). Velocity has to do with speed of data, volume with size and variety with variability (Zikopoulos and Eaton 2011). Big data is commonly used to analyse unstructured data (Suthaharan and Shan 2014). However, contrary to popular understanding, structured data can also be Big Data in as much as such data have the aforementioned characteristics (Zikopoulos and Eaton 2011). The most common and complete Big Data framework is Apache Hadoop.

Even though size is a vital feature to be assessed when trying to decide if a dataset qualifies as ‘Big Data’, the intensity of the computation required for the intended analysis on the dataset is important as much. This is most evident in Jacobs and Adam's (2009) research work where they fabricated data of the world’s population with focus on people’s demography (race, religion, income, employment, among others). This data was formulated in a table of more than 7 billion rows and about 10 columns, and was successfully stored on a 100 gigabyte hard disk. Modest commands written to answer simple questions like average height of world population were easily operational on a standard everyday computer hence, despite the large size of the data, it did not qualify as Big Data. Jacobs and Adam (2009) tried unsuccessfully mount the fabricated data on an enterprise-grade database system (PostgreSQL6) operating on a high performance computer (a workstation with 20 gigabyte RAM and two terabytes of hard disk). Despits having not attempted any analysis, the task had to be terminated after six hours. In the event of waiting for successful mounting, potential analyis would probably hav taken very many days, thus qualifying the same data as Big Data, based strictly on analysis type.

The given illustration is exactly why data of 100,000s of construction firms may or may not be taken as Big Data. A simple layout of such data on SPSS to find some mean averages will be easy and cannot be taken as ‘Big’. A more complicated analysis of the same data with highly tuned neural networks, for example, could take hours and qualify it as ‘Big Data’

1.4 Justification of study

The Department for Business Innovation and Skills (2013b) clearly stated that the construction sector is among the biggest sectors of the UK economy. The Department for Business Innovation and Skills (2013b) went further to explain that

> "*construction also has a much wider significance to the economy. It creates, builds and maintains the workplaces in which businesses operate and flourish, the economic infrastructure which keeps the nation connected, the homes in which people live and the schools and hospitals which provide the crucial services that society needs. A modern, competitive and efficient CI is essential to the UK's economic prosperity. Its contribution is also vital if the UK is to meet its Climate Change Act commitments and wider environmental and societal obligations*" (p. 2).

According to Rhodes (2015) in a House of Commons Library research paper, the CI in 2014 contributed £103 billion in economic output, representing 6.5% of the total; it also provided 2.1 million jobs or 6.2% of the UK total in 2015.

Despite the huge importance the CI has to the UK, the industry witnesses some of the highest rates of failures. In 2012 the construction sector insolvency rate in the UK was third highest at 14.4 percent (Dun and Bradstreet Limited, 2012). In England and Wales alone, construction businesses made up 23% of the total number of all businesses forced into compulsory liquidation in 2012 (Hodgson, 2013). Most recently, the industry again possessed the highest number of liquidated companies in the 12 months finishing in quarter two (Q2) of 2016 with a total of 2976 companies liquidated (The Insolvency service, 2016). This included 833 obligatory or forced liquidations and 2143 unforced liquidations (Figure 1.1).

Overall, the rate of construction business failure is unacceptable since it can lead to an economic downturn. Although efforts have been made to reduce the rate of failure, great success has not been achieved. Part of the efforts includes the development of insolvency prediction models (IPM) for multiple industries to aid early realisation of potential failure; this allows mitigation action to be taken early enough. The continuous failures, however, show that there is a need for robust IPMs to be developed specifically for the CI of a particular region so as to improve performance. The is because the construction industry is quite distinct from other industries (Bal, Cheung and Wu, 2013).

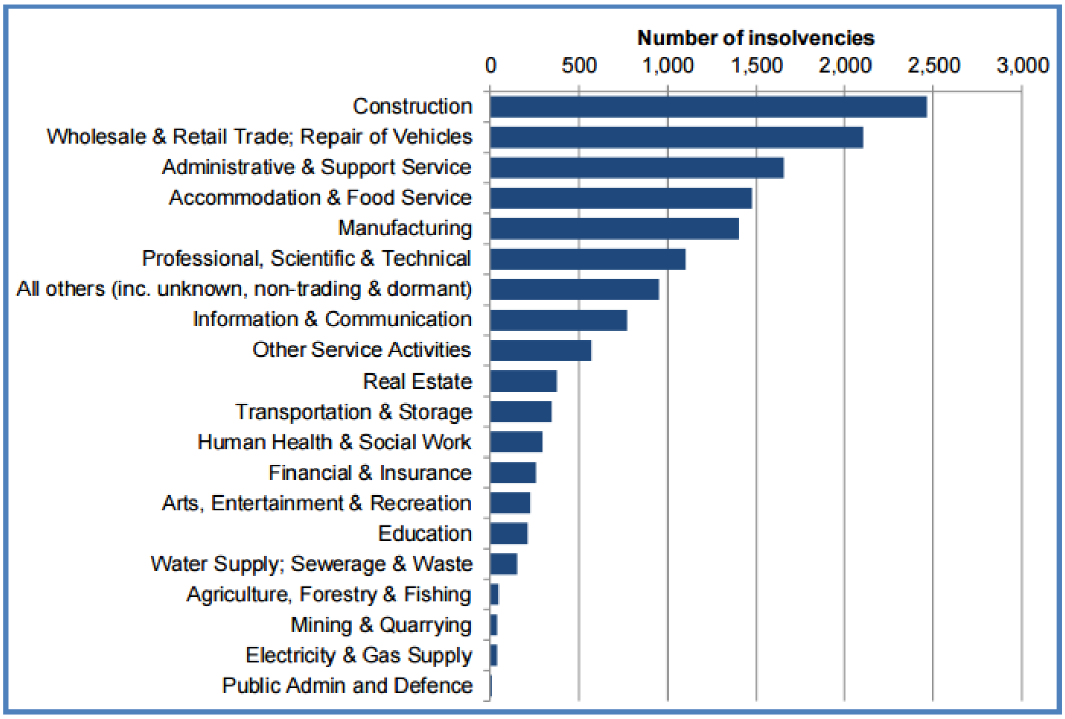


Figure 1.1: Total Company Liquidations in England and Wales by Broad Industry Sector, year ending 2013 Q2 (The Insolvency service, 2016)

1.5 Research problem and gap in knowledge

Most of the research conducted on developing IPMs are from finance professionals (e.g. Altman *et al.* 1994; Atiya 2001; Ko *et al.* 2001; Agarwal and Taffler 2008 and more), and were conducted to assess the creditworthiness of firms. Despite interest from owners and comparable stakeholders in preventing their firms from failing, only limited IPM studies (e.g. Ahn *et al.* 2000; Ko *et al.* 2001b) have made this their focus, none of which considered construction firms. Such IPMs take result transparency seriously because result interpretation is what allows the poor performance areas of a firm to be identified and given the required attention. Since the intention here was to help reduce failure of construction firms as explained in sections 1.1 and 1.4, transparency was a major requirement for the CF-IPM developed.

Only very few CF-IPM studies have used AI tools for ther models despite their very many benefits over statistical tools as proven by many studies (e.g. Yoon and Kwon 2010; Kim 2011; Huang *et al.* 2012; Wang *et al.* 2014). The adopted AI tools are the old but effective neural networks and support vector machine. None of the reffered studies had its focus on the UK construction industry (CI). Worst still, no CF-IPM study has adopted contemporary sophisticated AI tools like Bart Machine and Random Forest, leading to non-optimal models. Both old and contemporary AI tools were thus be used for model development in my research, with focus on the UK CI.

The only study to have attempted to use a relatively large data of 500 sample firms and a well-tuned AI tool (artificial neural networks in this case) ran into tough computation intensity problems. The effect of this was a very good model and tedious computational duration of five days with 30 PCs running Windows. Since large data increases reliability, and high tuning of AI tools lead to high performance, they were deemed important for the model to be built here. To solve the computational intensity problem, Big Data Analytics was used for model development since it could handle days' worth of computation in seconds.

Further, virtually all IPM studies have used only financial ratios (quantitative variables) to build models, neglecting the non-financial indicators usually in the form of qualitative variables. This is however not a very robust approach, as pointed out by many researchers (Argenti 1980; Zavgren 1983; Keasey, and Watson 1987; Kangari 1988; Hall 1994; Abidali and Harris 1995; Becchetti and Sierra 2003 among others), for at least three important reasons:

1) The exclusive use of financial ratios readily excludes firms with incomplete accounting data, a major feature of medium small and micro (MSM) construction firms. Also, some MSMs simply do not produce statements at all, and where MSMs have complete accounting data, they can easily be misrepresentations due to MSMs' common practice of outsourcing financial statement simply to meet legal requirements (Balcaen and Ooghe 2006). The exclusion of MSMs does not reflect the skewed distribution in the construction industry according to its statistics: the industry boasted over 950,000 small and medium enterprise (SME) in 2015; the industry represents circa 20% of the UK private sector SMEs, making it the sector with the highest percentage of SME firms (Department for Business Innovation and Skills, 2015); over 96% of UK construction

firms as of 2001 are small or micro firms (Jaunzens, 2001); and 86% of employees in the sector work in small construction firms (Jaunzens, 2001). These statistics show that insolvency in the construction industry cannot be reduced if MSMs are not included in the proposed solution. The CF-IPM to be built in my study will thus take MSM construction firms into consideration.

2) For a CF-IPM to have better early predictive capabilities, it needs to include qualitative variables which measure managerial decisions effect, company activities effect, the effect of personnel skill level, among others (Abidali and Harris 1995). This is because it is the result of these activities that translate into the numeral values of financial ratios. Early predictive capability of a CF-IPM is very important for firm owners and other stakeholders since the early prediction of potential failure allows more time for remedial actions to be taken (e.g. Hall 1994; Abidali and Harris 1995). Quantitative and qualitative variables will thus be combined for the CF-IPM to be developed in our proposed solution.

3) Finally, the CI is quite distinct from other industries (Bal, Cheung and Wu, 2013) and deserves to have industry specific models that reflect its activities. Although the numeric values range of financial ratios might vary by industry, the use of only financial ratios still makes models somewhat generic as financial reports are not industry specific. The use of qualitative variables on the other hand allows for industry specific events that will reflect the activities of the construction industry to be used. For example, measures for the effect of 'fluctuation of material cost', 'percentage of bids won', 'percentage of works completed to schedule' and the likes can and will be included as variables in the model to be developed as our proposed solution. This will guarantee that the model is customised to the CI and will avoid generic models as built in past studies.

Overall, it can be concluded that to solve the identified research problems, the CF-IPM to be built must:

1) be transperent enough to allow interpretability of result,
2) be chosen based on a comparison of models built with various powerful AI tools,
3) be built with relatively large data using Big Data Analytics,
4) be built with combined quantitative and qualitative variables.

1.6 Research questions

The research questions were formulated in consideration of the exposures in the 'Research Problem and Gap in Knowledge' section. They are as follows:

1. What qualitative variables contribute to solvency/insolvency of construction firms?
2. Which financial ratios are commonly reported by large, medium, small and micro construction firms?
3. Which are the best combined quantitative and qualitative variables for a CF-IPM?
4. How can dependability of the CF-IPM be assured in terms of tools and methods?
5. How can a highly reliable CF-IPM be developed with large data and tuned AI tools without running into computing troubles?

1.7 Aims, objectives and research questions

The aim is to develop a transparent holistic CF-IPM using a combination of qualitative and quantitative variables with the most sophisticated artificial intelligence tools mounted on a Big Data Analytics platform. The objectives that will help to achieve this aim and answer the research questions include:

1. To identify qualitative variables that contribute to solvency/insolvency of construction firms through literature review and fieldwork.
2. To identify the quantitative variables (i.e. financial ratios) that are commonly reported by large, medium, small and micro (LMSM) construction firms.
3. To select the best combination of quantitative and qualitative for the CF-IPM
4. To use advanced well-tuned AI tools and the best contemporary methods to ensure dependability of the CF-IPM

5. To solve the high computation intensity problem of large data and tuned AI tools by using Big Data Analytics to develop the CF-IPM.

1.8 Unit of analysis

Unit of analysis, according to Tainton (1990 p.5), "is the entity on which there are data and which will be subjected to analysis." It is the social unit about which data is collected; hypotheses are designed, and conclusions are made (Yang and Miller 2008). Although these explanations sound fairly straightforward, it is not uncommon to confuse what the actual unit of analysis in a study is. Grünbaum (2007; p. 82) in his research was able to explain/conclude that "the key issue in selecting and making decisions about appropriate unit of analysis is to decide what it is you want to be able to say something about at the end of the study". The unit of analysis is therefore absolutely dependent on the design of the study, and it is what viable conclusions about the study are mainly based on.

With the afore clarifications, the unit of analysis of the research was resolved to be construction firms, which falls under the 'organisations' category of unit of analysis. This is because the data collected and analysed for the research were those of construction firms. Also, the conclusions were made mainly on construction firms.

Although they are usually the same, the unit of analysis in a study might be different from the unit of observation which "is the entity on which the original measurements are made" (Tainton 1990, p.5). The original measurements in the research, e.g. financial ratios of construction firms, were made directly on construction firms hence the unit of observation was also construction firms.

1.9 Methodology

A major objective of the research was to involve qualitative variable in the IPM to be built for construction firms. There are a few studies (e.g. Kangari 1988; Hall 1994; Arditi *et al.* 2000) that have focused on factors that lead to insolvency of construction firms over the years. Some of the qualitative variables required can be deduced from these established factors. However, these studies will be missing some important contemporary dynamic factors. A very good example is the effect of immigration on the UK construction firms; is it positive or negative of large, medium, small and micro (LMSM) construction firms? There can also be issues with industry culture and the new sustainability policies. All these meant that it was good idea to talk to firm owners and senior management staff (respondents) of failed and existing construction firms to establish what contemporary factors have affected the solvency of their firms. This was done by interviewing respondents. The resulting factors were operationalized to form questionnaires. The result from the questionnaires (qualitative variables) and the financial ratios (quantitative variables) of sample firms were used together to create the CF-IPMs. This approach of using any combination of methods that best answer the research method, irrespective of the school of taught, is known as pragmatism (Tashakkori and Teddlie 1998). The method used was mixed method which normally ensures an all-round effectiveness research (Creswell and Plano Clark 2011). The sample, data collection and analysis methods are briefly expatiated on in the following paragraphs.

Sample: The sample population was of failed and existing construction firms of all sizes in the UK. The failed construction firms considered were those that failed between 2009 (after the global recession) and 2016. To have a sample that is more representative of the population, medium, small and micro (MSM) construction firms constituted 'at least' 80% of the sample with large firms forming the rest. Although there are normally a lot more existing firms than failed firms, the variance in the proportion of failed to existing firms in the data was not allowed beyond a ratio of 6:4 to avoid variance problems in analysis. The sampling methods are discussed in chapter six.

Literature Review: A systematic review of studies that have identified factors that lead to failure of construction firms, and not necessarily developed a CF-IPM, was carried out. The factors deduced from the review were operationalized, along with factors realised from interviews, and used to create a questionnaire.

Unstructured Interviews: The qualitative data was gotten using interviews. "The unstructured interviews take the form of free-flowing conversation" and are known for the advantage of not limiting respondent views (Latham and Finnegan 1993; p. 42). The unstructured interviews were conducted orally with owners and senior management staff of failed (or insolvent) and existing construction firms of various sizes. The interviewees were simply asked to talk freely about what they think had contributed to the failure or survival of their construction firm. Every other question was generated from the responses given. The responses were recorded and analysed. The insolvency factors gathered from the unstructured interviews and literature review were subsequently operationalized and used to develop questionnaires.

Questionnaires: The questionnaires developed from operationalizing factors gotten from literature review and interviews were distributed to a much larger number of respondents (i.e. owners and senior management staff of failed and existing construction firms). To encourage participation of target respondents, Likert scale questionnaires with closed-ended questions were used because Likert scale questionnaires are quite easy to deal with by respondents (Van Laerhoven, van der Zaag-Loonen and Derkx, 2004). Questionnaires were distributed and collected mostly by post, using prepaid envelopes, and by email where possible. The response from the questionnaires were used as qualitative variables' values in building the CF-IPM.

Company Documentation: Data for financial ratios of sample firms were downloaded from FAME (Forecasting Analysis and Modelling Environment) Bureau Van Dijk UK financial database. This database is available for free on the university system. The data downloaded were those of the firms whose representative completed the questionnaires. Only the financial ratios from the last year of operation, or year of failure where applicable, were used. Where the financial statement of last year of operation was not available, that of the preceding year was used. Where the financial statement of sample construction firm had very scanty information, the financial statement of a firm that was similar in features (i.e. size, turnover, number of employees, profit, date of establishment, insolvency date where applicable and location) was used as a replacement

Analysis of Data: For analysis, the data collected was be split into two sets: 70 percent for model development and 30% for model testing. The proportion of failed to non-failed firms in each split was (approximately) the same. The quantitative and qualitative variables, as measured for all sample firms, were analysed using various statistical tests like 'information gain' to find the variables that best distinguished between failed and existing construction firms. The selected variables were subsequently used to build CF-IPMs using various powerful AI tools (e.g. AdaBoost, support vector machine, random forest, Bart machine, among others.) and the results compared to find out the best model.

1.10 Novelty of research

With too many studies somewhat re-inventing the wheel, the importance of novelty in contemporary research is becoming more important than ever. Many studies have shown that novel research contributes significantly to the economy; this is very necessary as the research world itself consumes huge funds and needs to feed back into the economy somehow (Griliches, 1979; Etzkowitz *et al.*, 2000).

1.10.1 Academic and literature novelty

The research work aspired to develop IPMs for early identification of potential failure of construction firms and was novel in a number of ways. It was the first time that the combination of quantitative and qualitative factors (variables) that are the best predictors of solvency of construction firms are identified together. It was also the first time mixed variables (qualitative and quantitative) were used to develop CF-IPMs, thereby being the first case to establish a mode of combining them in a CF-IPM. Further, it was the first time to consider, simultaneously, the effect of managerial decisions and construction industry or construction firms-specific activities on predicting potential failure of construction firms.

The first systematic review of how tools perform in relation to CF-IPM development was carried out in the research work since no study has taken a holistic approach towards this area. The review exercise was used to identify what criteria each tool satisfied best and consequently the most fitting tools for various situations. Also, the best tool for predicting insolvency of LMSM construction firms was, for the first time, revealed in the research

1.10.2 Practice novelty

The Construction sector has always been, and remains, one of the most important sectors of any country. Apart from its huge economic importance, its role in providing and maintaining homes, vital infrastructures, schools, hospitals, and so forth means that large failure of construction businesses is always a national concern. In recognition of this and in a pioneering effort, the development of CF-IPMs with early warning system for construction firm owners and other stakeholders who are interested in detecting potential failure early enough, to allow enough time for possible recovery, was implemented in the research work. This was done mainly by including qualitative variables in the development of the model. This differed from CF-IPM developed in other studies that used only quantitative variables since those CF-IPMs were meant for credit providers who are interested in checking credit worthiness of construction firms. Consequently, the model developed will help stem the tide of the relatively massive failures of construction firms

The CF-IPM developed was practically novel in that it was the first to be developed with the data of large, medium, small and micro (LMSM), thereby being widely applicable and not neglecting the MSM construction firms that make up about 90% of the UK construction sector as done in previous studies.

1.11 Scope and limitations

The scope of any research work is very important to know how widely generalizable the results is. The scope here is discussed mainly in two dimensions: type and size of construction firms.

Regarding size, the scope of this work included all sizes of construction firms in the United Kingdom i.e. large, medium, small and micro. However, to have a sample that is more representative of the population, medium, small and micro firms constituted ‘at least’ 80% of the sample with large firms forming the rest. The definition of firm sizes according to the European Union is given in Table 1.3

Table 1.1: Categories of firm sizes according to the European Union

Company category	Staff headcount	Turnover	or	Balance sheet total
Large	250 and above	≥ € 50 m		≥ € 43 m
Medium-sized	< 250	≤ € 50 m		≤ € 43 m
Small	< 50	≤ € 10 m		≤ € 10 m
Micro	< 10	≤ € 2 m		≤ € 2 m

Regarding firm types, the construction firms considered in the research were those that are involved directly in on-site constructions. These are the ones classified by the UK Standard industrial classification of economic activities (SIC) 2007 as 41100 Development of building projects; 41201 Construction of commercial buildings; 42110 Construction of roads and motorways; 42120 Construction of railways and underground railways; 41202 Construction of domestic buildings; 42130 Construction of bridges and tunnels; 42210 Construction of utility projects for fluids; 42220 Construction of utility projects for electricity and telecommunications; 42910 Construction of water projects; 42990 Construction of other civil engineering projects n.e.c.; 43110 Demolition; and 43120 Site preparation. It does not involve 43130 Test drilling and boring; 43210 Electrical installations; 43220 Plumbing, heat and air-conditioning installation; 43290 Other construction installation; 43310 Plastering; 43320 Joinery installation, etc.

The main limitation of this work was the inability to recognise construction firms that declared false bankruptcy and got away with it legally. Such firms, having successfully declared bankruptcy under the law, could be wrongly included in the failed construction firms sample population. If such a firm is chosen as a sample construction firm, its attributes would be wrongly processed by an IPM tool with those of other failed firms, thereby reducing accuracy. It must, however, be mentioned that any construction firm with suspected false bankruptcy declaration was excluded from the samples to be used for the research. A relatively less critical limitation was the inability to interview all willing representative of construction firms who insisted on physical presence despite being located far from the author, because of the cost barrier. This reduced the number of interviews held. Another relatively less critical limitation was the consideration of construction firms that have failed over a seven-year period, i.e. from 2009 to 2016 since it might mean the firms considered were subjected to different external factors that led to their failure. However, the prediction tools can still find a pattern amongst these failures and ensure this limitation is reduced to the barest minimum.

1.12 Thesis structure

The structure of this thesis flows from this introductory chapter through a series of literature reviews touching on underpinning theories, variable influencing CF-IPMs, Big Data and methodical issues with developing CF-IPMs. These reviews partly involved systematic reviews, some of which were subsequently used to decide the methodology of the thesis. Data from methodology was processed using analyses methods like Cronbach alpha reliability and factor analysis to create the variables to develop the CF-IPMs. The CF-IPMs were subsequently developed using Big Data Analytics and sophisticated AI tools. The results from the analyses were finally discussed and conclusions made. A framework of the thesis structure is provided in Figure 1.2.

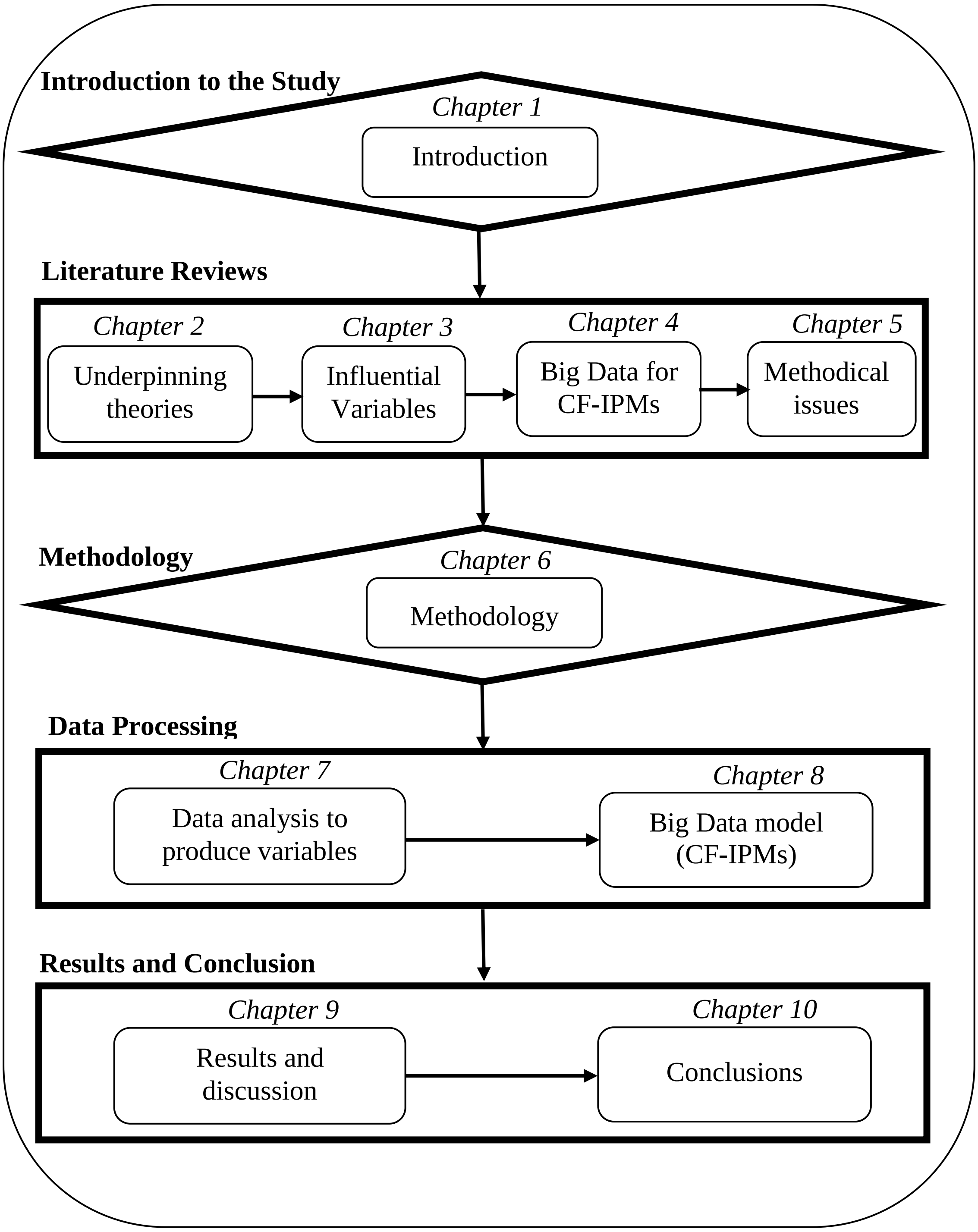


Figure 1.2: A framework of this thesis structure

1.13 Key Achievements

A major achievement in my research work was the successful development of a CF-IPM that is relevant for all sizes of construction firms: large, medium, small and micro. Despite its necessity (90% of the construction industry is made up of medium, small and micro firms), this feat has neither been attempted nor accidentally achieved by past studies. This is because past studies focused solely on large construction firms because they are thought to be more important.

Another achievement was the successful use of combined qualitative and quantitative variables to develop a CF-IPM. The closest achievement to this was witnessed in Abidali and Harris' (1995) study where two separate CF-IPMs were developed, one with quantitative variables and the other with qualitative.

A major accomplishment was the successful use of Big Data Analytics (BDA) to develop CF-IPMs with relatively large data and highly tuned AI tools. The development of this type of CF-IPM would normally require days to complete due to the high computation intensity that results from the data and highly tuned AI tools. In this case however, the BDA platform helped ensure that the models development was completed in minutes, or even seconds in some cases.

Finally, the use of contemporary powerful AI tools helped achieve CF-IPMS with over 99% accuracy, a rare feat among CF-IPM studies.

1.14 Chapter summary

This chapter gave an overview of the problem of high rate of construction firms failure in the UK and Europe and the effect of such failures on the economy. It explained the concept of failure and introduced the concept of Big Data. The relatively high rate of failure of construction firms was used to justify the research. The research problem and gaps in knowledge exposed the poor over-reliance on financial indicators of past construction industry IPM studies and their neglect of micro and SME construction firms which constitute over 90% of the CI. It also explains how the studies have failed to prioritise early predictive capabilities of IPMs which firm owners need to allow time for remedial actions and in turn reduce the rate of firm failure in the CI. The proposed solution thus offers to combine financial (quantitative) and non-financial (qualitative) indicators (variables) to develop a robust IPM to help reduce the rate of construction firms’ insolvency. The proposed solution further offers to use large data set for reliability sake and state of the art tools, a combination which calls for the use of the contemporary Big Data Analytics, to develop the CF-IPMs.

The unit of analysis was cleared up as being construction firms (failed and existing) which correspond to the ‘organisation’ category. In order avoid confusion during sample selection, the concept of failure is clearly explained and what failure refers to in the research is spelt out. A brief methodology section was used to highlight how pragmatism is the philosophical underpinning of the research and justify the mixed method approach used and a thesis structure provided

The novelty of research section highlighted that the research is the first to combine qualitative and quantitative variables, use Big Data Analytics, and use the data of all sizes of construction firms (i.e. LMSM) to develop a holistic CF-IPM. Finally, the scope and limitation section defined the scope of the work according to construction firm types and sizes. A structure of the thesis was also given.

Chapter two contains a discussion of the theories underpinning failure of construction firms. The theories were discussed from three major dimensions: external based theories, internal based theories and mixed (or combinatory) theories.

CHAPTER TWO

2.0 UNDERPINNING THEORIES OF CONSTRUCTION FIRMS FAILURE

2.1 Chapter introduction

Construction firm insolvency research has been approached principally from the accounting and finance standpoint by developing insolvency prediction models using financial ratios exclusively (e.g. Baum and Singh, 1996; Fadel, 1977; Horta and Camanho, 2013; Langford, Iyagba, and Komba, 1993; Mason and Harris, 1979; among others). For a wider, more rigorous approach, there is need to consider insolvency of construction firms from non-financial theoretical perspectives such as organisational theories. This chapter introduces these theories and relates them to the failure of construction firms.

The next section discusses organisational theory categories (external, internal and combinatorial or mixed) that are considered in the research. Section 2.3 is about external based theories, including organisation ecology and Porter's perspective which is explained in subsections 2.3.1 and 2.3.2 respectively. Section 2.4 is about internal based theories, including the adaptationist perspective, Mintzberg's perspective, upper echelon theory and resource based view which are explained in subsections 2.4.1, 2.4.2, 2.4.3 and 2.4.4 respectively. Section 2.5 is a description of dynamic capabilities theory which is the mixed theory considered in the research. Section 2.6 is a highlight of the implications of the discussed theories. Section 2.7 provides a summary of the chapter

2.2 Theory categories

Organisational theory has to do with how the analysis of organisations are done in the sense of identification of difficulties and their solutions, and maximising efficiency, effectiveness and performance. It has to do with the structures put in place in an organisation and how an organisation is designed to function. It deals with internal and external, or even mixed (internal and external), relationships of an organisation (Figure 2.1). Regarding construction firms, it deals with how the firm relates with its clients, suppliers, subcontractors, and the likes, as well as how the firm relates with its employees right from the director level to the junior engineers' level.

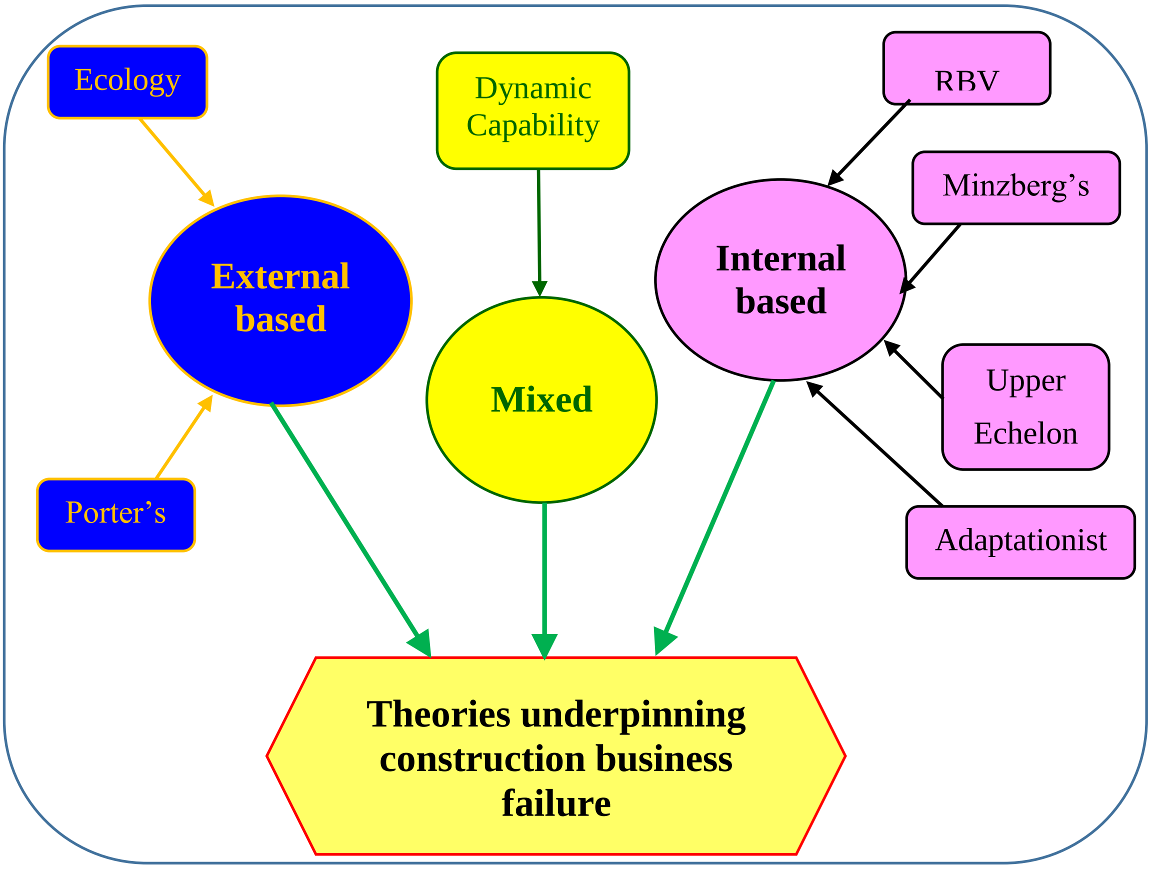


Figure 2.1: Structure of theories underpinning construction firm failure

Given the multifaceted nature of insolvency and that no individual criteria can answer what determines insolvency of a firm, there are numerous contending theories attempting to reveal what helps to improve solvency (i.e. to avoid failure or insolvency). Many theories on what strategy to use to aid solvency have been developed over the years. The theories are quite variant and are not necessarily mutually exclusive, most of them having a different emphasis. The diverse and complex nature of strategy was attested to by Mintzberg, Ahlstrand, and Lampel (1998). After reviewing ten strategy models, they concluded that "strategy formation is judgmental designing, intuitive visioning, and emergent learning; it is about transformation as well as perpetuation; it must involve individual cognition and social interaction, cooperation as well as conflict; it has to include analysing before and programming after as well as negotiating during; and all of this must be in response to what can be a demanding environment. Just try and leave any of this out and see what happens" (Mintzberg *et al.,* 1998: pp 372-373). The following sections describe some applicable theories to the research

2.3 External based theories

2.3.1 Organization ecology

The organisational ecology perspective attempts to encompass the complete variety and diversity of firms through their beginning, development, change period, and death. According to George (2002), Hannah used the word ecology because he stumbled on a population ecology document while looking for conceptual models to support his ideology. Right from early days of organisational ecology, the very presence and different destinies of various firms that failed and survived mattered, none being more important than the other. Pioneering research in the area looked at entire populations irrespective of size or duration of firm existence. The smallest and biggest firms, short-lived and the very oldest firms, all are important in organisation ecology. Hannan and Freeman (1977) (Hannah in particular), the fathers of organisation ecology, conceptualised the idea in the form of rebuttal to organisation research's focus on large firms. Hannah advocated the recognition of thousands of small firms that were unnoticed as data involving them could be key to solving organisational problems.

More importantly, Hannan and Freeman (1977) rejected the idea that organisations do transform or adapt and questioned why many firms were failing if they could adapt. The idea, more or less, is similar to Darwin theory of natural selection. This theory postulates that the environment selects firms whose traits/characteristics are most fitting for survival while firms that have traits/characteristics that are not fitting to the environment will fail and get replaced by new ones (Kale and Arditi 1999).

While arguing that all firms in a population are important and that all types of firms should be considered, proponents of organisational ecology contend that the population include firms that were considerably planned for but eventually did not start up. Some of these planning activities normally result into successful start-ups while others result in abandonment or a 'dead on arrival' situation. Hannan and Freeman (1977) argue that a lot of time and resources are put into the said planning that the firms cannot be disregarded as part of the population. Disregarding these unsuccessfully started firms, in ecology as postulated by Hannah, leads to a serious underestimation of failed firms. Carroll and Hannan, (1995) successfully collected such data (covering from 1886 to 1994) in the automobile industry in an investigation in 1994 and realised that close to 4000 potential automobile production firms did not eventually get to production, meaning that about 89% of potential automobile firms failed to reach operation stage. As noted in George's (2002) article, collecting this sort of data is onerous and distressing and consumes hundreds of hours. This is partly why this part of organisational theory will be adopted in the research, especially as it is nearly impossible to do this with construction firms i.e. collect data of all unsuccessful start-up of construction firms.

Organisational ecology's take on competition argues that increase in population lead to legitimation and competition (Hannan and Freeman 1977). Legitimation is when a particular method of activity execution becomes the norm in the industry, and this leads to the birth of more firms. Competition has to do with when the number of firms is so much that the available resources, including customers, become insufficient (Amburgey and Rao 1996). This is a case of market saturation which is very common in the construction industry.

2.3.2 Porter's perspective

The Porter's perspective is famous for the five competitive forces model: supplier power, buyer power, competitive rivalry, the threat of substitution and threat of new entrants. According to (Rumelt, Schendel, and Teece, 1991: p.8), "the most influential contribution of the decade from economics was undoubtedly Porter's competitive strategy (1980)". It has been the basis for some strategy research in construction (e.g. Betts and Ofori, 1992; Budayan, Dikmen, and Talat Birgonul, 2013; Tansey, Spillane, and Meng, 2014). The threat of new entrant remains one of the most applicable forces to the construction industry as the entrance to the industry has no barrier and sometimes require little investment (Betts and

Ofori 1992). This is unlike some other industries like the computing and engineering industries in Japan where huge investments by larger companies are proving to be a barrier to entrance for potential smaller companies. Supplier power wise, there are usually many suppliers in the construction industry, however, keeping a good relationship with a small set of specific suppliers, thereby buying in high volumes from them could give a competitive advantage. This is because being a major buyer allows the firm to drive down prices of the supplier. It also ensures the firm is given priority when there is materials shortage. The threat of substitution refers to how easy it is for a client to replace one firm with the other. This threat is usually high in the construction industry as there are always too many firms competing for one job, hence being unique can give a competitive advantage here. Competitive rivalry, which is the fifth force is all about firms vying for a better/unique position to give them a competitive advantage. According to Betts and Ofori (1992), vying for position is a strong competitive force among small construction firms despite the low exit barrier of the industry.

2.4 Internal based theories theory

2.4.1 Adaptationist perspective

The adaptationist perspective (Thompson, 2014; Child, 1972; March, 1981; Bourgeois, 1984) is quite the opposite of organisational ecology in that it accepts that firms can change and adapt to the environment to survive. This is in line with Lamarck's ideology, known as Larmakism, which stresses how important the adaptation aspect of an organisation is. As opposed to ecology research which prefers to consider the full population, research in the adaptationist world usually considers just one organisation (Kale and Arditi 1999). Adaptation in this sense has to do with the selections, pronouncements and general activities that the top management (or owner in the case of micro firms) of an organisation make to ensure it fits properly to its environment. Take for example, many construction firms in the United Kingdom (UK) are taking the steps of improving their expertise in the Building Information Modelling (BIM) area because the UK government activated the BIM mandate in April 2016. This mandate means any firm that cannot operate at BIM level 2 cannot get a government contract. These steps have seen many firms change the way they do some things including documentation, construction methods and recruitment, among others. This sort of situation is clearly a case of adapting to a changing environment to ensure a better chance of survival than failure, especially if a firm in the case of the given example is very dependent on government projects for survival.

Theories of Competition: The competition aspect of organisational theory has a take on Red Queen theory in which ecology and adaptationist perspectives cross path. The theory suggests that improvement of some firms in a competitive market is as a result of deteriorating conditions for others who must adapt to the climate of an evolved environment to survive (Barnett and McKendrick 2004). The adaptation comes in the form of some organisational learning as the deteriorating firms (or organisations) continue to understand the new environment within which they operate. Successfully adapted firms that were deteriorating, in turn, become upgraded and put pressure on the previously improved firms and vice versa, therefore causing an interminable circle of organisational learning and competition (Baum and Singh 1996). This leads to very fierce competition, as witnessed in the construction sector where bid writing, for example, has been learned so much by most firms. These firms are now able to submit an incomprehensibly low bid for jobs simply because of their expert knowledge of where extra savings or future charges to the client can be legally made. The competition is so fierce that firms use unrealistic tenders to win projects that consequently lead to failure of the firms (Arditi, Koksal and Kale, 2000).

The Red Queen theory was well supported in Barnett and McKendrick's (2004) study where they were able to prove that firms that are very engaged in thick competition over time learn a lot and become stronger with a very reduced potential for failure. On the other hand, the avoidance of competition either by focusing on a particular region or on a particular area of the industry, although has its immediate benefits, is a less robust strategy that will hardly bring about learning and associated innovative knowledge. Engagement in thick competition in itself can be as a result of reduced availability of construction projects (Kangari, 1988). Although Kangari (1988) did not disagree with the organisational learning that comes with fierce competition, he clearly highlighted that a reduction in available construction projects and the resulting fierce competition leads to failure of construction firms.

Organisational learning and liability of newness: Organizational learning obviously has to do with age as learning takes place over time. Together with size, they have been given the most attention in research. The fact that most start-up firms are of small size means it is almost impossible to separate the two of them according to Wholey and Brittain, (1986) but authors like Ranger-Moore (1997) argue that age, which is associated with the liability of newness hypothesis, is more important than size.

Liability of newness explains how external and internal factors lead to (in)solvency of new firms. Internal factors have to do with functions like organisational structure, positions in the structure, personnel skills and experience, relationship types and level among staff, among others (Freeman, Carroll and Hannan, 1983). Over time, these functions are improved upon by the company due to experience/organisational learning (Crossan, Lane and White, 1999). The firms that fail to learn and adapt thus end up failing. Although Carroll and Hannan (1995), talking from the ecology perspective, disagree with this concept by claiming data that supports this stance is heavily biased. They (i.e. Carroll and Hannan, 1995) insist that most firms will normally fail, rather than adapt, in a changing environment and be replaced by new firms that are suited to the new environment hence the portended age effect is actually a size effect. Hannan and Freeman's (1977) earlier study, however, agrees that a firm learns how to do things in a more structured way over time, thereby improving its reliability, and this is the basis for the theory of `structural inertia' according to (Levinthal and March 1993). The theory postulates that reliability, which stems from having an explicitly stated repeatable structured way of doing things due to organisational learning, leads to improved solvency of a firm.

The external factors have to do mainly with an organisation's external relations such as clients, sub-contractors, material suppliers, financiers, among others. Bruderl and Schussler (1990) argue that developing a very good reputation with external relations, through meeting or exceeding their expectations in prior projects, remains very key to the survival of an organisation. This process will usually put an organisation on the top of the list when there is competition for resources from the external relations.

2.4.2 Mintzberg's perspective

The Mintzberg's perspective is famous as it took a holistic and integrated approach to various strategy theories to develop what is known as the five P's (plan, ploy, pattern, position and perspective). It has been the main or part basis for some strategy research in construction (e.g. Chinowsky and Meredith, 2000; Dikmen and Birgönül, 2003). A strategy is more or less in itself regarded as a plan. Planning is the most popular and is virtually the default approach by managers. It is usually based on information hence having poor information can lead to poor strategy as plan. Ploy strategy mainly has to do with making a ploy to outwit competitors while the pattern is about the decision a firm takes over time which then becomes the firm's way of doing things (Adams and Simon 1962). According to Mintzberg *et al.* (1998), it is the actions that a firm takes, and not the decisions, that lead to patterns; this is because the interconnection between decision making and actions in a firm is usually unclear. There is often a great deal of action with little decisions, and sometimes vice versa. Further, the actions and decisions are sometimes uncorrelated. According to Andrews (1987), strategy as position refers to positioning a firm in such a way that it stands out from others. This is very much about being unique. Perspective as strategy refers to the fact that the ways of thinking in a firm will largely influence the strategy the firm adopts. For example, a firm that encourages caution in resource consumption and waste generation is likely to have employees come up with more sustainable solutions.

2.4.3 Upper echelon theory

Upper echelon theory is a behavioural theory that suggests that the failure or existence of a firm partly has to do with those that constitute the top management (Huber and Glick 1993). The theory was proposed by Hambrick and Mason (1984) who reconciled preceding individual literature on how characteristics of the top management team partly affects the strategic fate of a firm. The top management team is usually defined in upper echelon theory as the people holding executive managerial position and member of the board of directors position simultaneously (Finkelstein and Hambrick 1990). Although many upper echelon theory applications have determined those involved in the top management team based on convenience (Carpenter, Geletkanycz and Sanders, 2004). In upper echelon theory research studies, the top management team is usually taken as people occupying the topmost managerial positions like Managing Director, Chairman Board of Directors, and the likes, and is details are usually taken from publicly available information.

The upper echelon theory argues that the knowledge, experience, education, social background, values, personality traits, among others, of top management teams influence the performance of a firm they lead. In essence, the outcome or potential status of a firm (whether it will fail or survive) can be predicted from top management team characteristics. According to Hambrick and Mason (1984), it is vital to consider the causal factors of the characteristics of the top management team when dealing with upper echelon theory. For example, the decisions and actions of members of the top management team that possess vast experience are more likely to be based on their experiences rather than their traits.

In the United Kingdom construction industry, where a large proportion of the firms are micro, small or medium in size (Department for Business Innovation and Skills, 2015), it is not uncommon for the top management team to comprise of just the owner of the firm. Hall (1994) conducted a study on insolvency of small and micro construction firms and is one of the very few studies to have employed the upper echelon theory in this field. The nature of his study sample, i.e. small and micro construction firms, meant that the top management teams were simply the owners of the sample firms. The study used a questionnaire that asked about the age, qualification, management experience, professional training, professional membership, among others of the owners. In the result, age at which owner took charge of the construction firm and education level of owners turned out to be very important variables in determining (in)solvency of a construction firm according to a regression model. This clearly proves the potential applicability of the theory to the construction industry.

Abatecola and Cristofaro (2016) tested the theory on ‘important’ construction firms in Italy with a focus on top management team members such as president, chief of operation, and partners. The construction industry was viewed from the value chain perspective by the authors hence it included preconstruction, post construction and other supporting actions. The result indicated similarity in the actions taken by members of the top management team with similar characteristics. In this regard, educational background, education level, age and gender were all found to be significant, moderately indicating agreement with the result of Hall's (1994) study. This could mean the effect of upper echelon theory in construction firms is insensitive to size, a feature which favours part of the research's objective of creating one model for big and small firms.

2.5.4 Resource based view

As opposed to external based theories like Porter's five forces that focus on the industry rather than the firm's ability, the resource based view is internal based and focuses on the resources of a firm. Resources in this sense include tangible and non-tangible resources. Tangible resources of a construction firm include, for example, owning facility for a special construction method (e.g. volumetric construction), owning special equipment (e.g. drones capable of measuring aggregate volume), owning contemporary software (e.g. BIM compatible software like AutoCAD, ArchiCAD, Navisworks, among others), among others. Non-tangible resources include the skills, knowledge and experience of the firm's personnel, especially at managerial level. Non-tangible resources of construction firms, for example, include industry knowledge of the TMT, ability to use building information modelling (BIM) to execute projects, among others

The resource based view portends that the resources and proficiencies of each firm are unique and the survival of a firm is based on these resources. The theory postulates that a firm's resources are the main source of its competitive advantage. In fact, Wernerfelt (1984) explained that a firm is best looked at as a gummy assortment of resources. This gummy assortment of resources that defines a firm makes each firm unique and hard to replicate. Although one, two or more resources might be replicable, the holistic combination of the gluey assortment of resources is not. These resources can be rearranged for better performance and improved competitive advantage.

Not each and every resource of a firm is unique to the firm and creates a competitive advantage. According to Barney (2000), the resources that will give a firm an edge in the competitive market must be valuable, unique, have no alternative and not be perfectly imitable. According to Olavarrieta and Ellinger (1997), the more a firm uses its resources, especially the intangible ones like skills, the more it gets polished up, subtle and advanced it becomes. This continuous advancement and subtleness make the resource more unique to the firm and harder to replicate by other firms. One of many studies that have applied the resource based view theory to construction firms is Jaafar and Abdul-Aziz (2005). In the non-product based construction industry, uniqueness is usually about the method of execution, and this is normally dependent on the resources at the disposal of the firm (Korn and Pine 2014).

2.5 Mixed or combinatory theory

2.5.1 Dynamic capabilities theory

Dynamic capabilities theory focuses on the dynamism of a firm regarding the use of its resources. It supports, but goes beyond the resource based view. Dynamic capabilities theory portends that a firm' survival depends on in its ability to acclimatise its resources to the market demands to gain competitive advantage. According to Teece, Pisano, and Shuen (1997), dynamic capabilities is defined as a "firm's ability to integrate, build, and reconfigure internal and external competences to address rapidly changing environments" (p.516).

The dynamic capabilities theory heralds that core capabilities can be used to alter temporary competitive posture which can be used to develop lasting competitive advantage. So as to tackle fresh challenges in a constantly changing market, some dynamic capabilities essential. One is the learning stage where the firm and its personnel need to learn very fast and develop strategic resources. Another is that new resources such as technology (e.g. model integration software like Navisworks), capability (e.g. ability to use BIM for construction operations), among others, must be assimilated into the firm's system. And finally, the existing firm's resources must be converted and remodelled.

The dynamic capabilities theory according to Teece's perspective heralds that the important thing is for firms to possess corporate deftness, especially from three angles. Firstly, firms need to have a high level of awareness and should be able to recognise opportunities and dangers. Secondly, firms must be able to take the opportunities and avoid the dangers. Thirdly, firms must be continually competitive by continually advancing, integrating, conserving and, if crucial, converting its tangible and intangible resources.

2.5.2 Organisational co-evolution

The organisational co-evolution theory is more or less a combination of the organisation ecology theory and the adaptationist perspective. It basically postulates that organizations evolve based on the evolution of other organizations, and this co-evolution contributes to the business environment evolution. In essence, the organizations do not just evolve based on the business environment, rather they evolve based on one other thereby causing a change (evolution) in the business environment. Organizational evolution and business environment evolution thus happen simultaneously. According to Lewin *et al.*'s (1999) perspective, organisational co-evolution theory postulates that the organization population, the business environment and an organization itself are all associated consequence of managerial actions

This implies that strategies implemented through managerial actions are not simply lifeless reaction, instead they are a gung ho aim to transform both an organization and its environment. In the case of direct evolution, a pair of organizations go through evolution in reaction to each other while in the case of diffused evolution, one or many organisations go through evolution in reaction to numerous other organizations in the same business environment (Baum and Singh 1994). The organizational co-evolution theory has not been applied to the construction industry in research. According to Baum and Singh (1994), the theory's potential for application lies with very complex situations. It was however discussed here because of its connections to the afore discussed organization ecology (external based) and adaptationist perspective (internal based).

2.6 Implication of theories on construction firm failures

The discussed theories have various implications for construction firm failure (see Table 2.1). The organisation ecology causes construction firms top management team (TMT) to focus on its processes or methods and try to be unique such that it can withstand market problems. Porter’s perspective provokes construction firms TMT to look at everything from a relative perspective as it seeks to have a better competitive advantage. Such TMT will seek to have a relatively better relationship with its client and suppliers among others, and possibly submit lower bids than its competitors.

Table 2.1: Implication of Theories on Construction Firm failures

Underpinning Theory	Implication on failure of construction firms
External based	
Ecology	Will make Cause construction firms attempt to be unique to ensure environmental conditions does not lead to failure
Porter’s perspective	Cause construction firms TMT to view things relatively and improve relationship with stakeholders to improve solvency
Internal based	
Adaptationist perspective	Construction firms TMT do self-study and make decisions based on experience to adapt better to the industry and avoid failure.
Mintzberg’s perspective	Construction firms TMT do more planning and aim to stand out to get the competitive edge
Upper echelon	Construction firms TMT consider characters of people to be promoted/recruited to TMT level as characters will reflect decisive decisions that can decide the fate of firm.

Underpinning Theory	Implication on failure of construction firms
Resource based view	Construction firms TMT focuses on getting the right resources to improve competitive advantage
Mixed or combinatorial	
Dynamic capability	Construction firms TMT instigates continual resource upgrade

The adaptationist perspective causes construction firms' TMT to take more actions based on experience as the firm adapts to the construction industry. The Mintzberg's perspective provokes construction firms TMT to focus more on their own plans through which they try to achieve uniqueness and outwit competitors. Upper echelon theory leads to a construction firm's TMT considering characters of people before they are promoted/recruited to join the TMT as their characters will reflect the decisions they take, which can ultimately decide the fate of the firm. Resource based view causes construction firms TMT to focus more effort on recruiting highly skilled and experienced people as well as procuring cutting edge contemporary equipment. The dynamic capabilities theory has more of a mixed effect in that it supports the resource based view and adaptationist perspective. It motivates a construction firm's TMT to want to constantly upgrade/change the firm's resources to meet contemporary demands in the construction market

2.7 Chapter summary

This chapter introduced and discussed some non-financial organisation theories in view of achieving one of this studies main objectives of using qualitative variables (together with quantitative variables). The three categories of theory discussed are external, internal and combinatorial or mixed. The external based theories include organisation ecology and Porter's perspective. Organisation ecology suggests that the operating environment naturally selects the firms that will survive and leave those that will fail. Porter's perspective is famous for the five competitive forces model: supplier power, buyer power, competitive rivalry, the threat of substitution and threat of new entrants.

The internal based theories include adaptationist perspective, Mintzberg’s perspective, upper echelon theory and resource based view. Adaptationist perspective contends that organisations do learn, improve and adapt over time, leading to fierce competition that causes failure of firms. The Mintzberg’s strategic theory considered in the research is the five P’s: plan, ploy, pattern, position and perspective. The upper echelon theory explains how the characters of TMT members affect decisions that are key to the survival of a firm. Resource based view portends that the resources of each firm makes it unique and decides if it will fail or survive.

The mixed or combinatorial theories discussed are the dynamic capabilities theory and organisational co-evolution. Dynamic capabilities theory has a mixed effect in that it supports the resource based view and adaptationist perspective at the same time while organisational co-evolution supports both organisational ecology and adaptationist perspective.

Literature on the qualitative and quantitative variables that have been used in developing construction firms insolvency prediction models were reviewed in chapter three. A review of qualitative variables affecting general insolvency of construction firms was also included.

CHAPTER THREE

3.0 QUANTITATIVE AND QUALITATIVE VARIABLES INFLUENCING CONSTRUCTION FIRMS INSOLVENCY PREDICTION MODELS

3.1 Chapter introduction

The two key issues that most affect the performance of a construction firms insolvency prediction model (CF-IPM) are the variables and methods used. Variables are extremely important as they are the features that decide if an insolvency prediction model is for construction firms or not. One of the critical arguments on CF-IPMs is the relevance of the variables used to the construction firms because of the construction industry's uniqueness. The vital importance of qualitative variables, as against qualitative variables, as the variables that can really reflect the construction industry features necessary to build a valid CF-IPM have long been advocated and are supported here.

This chapter is thus geared towards a review of the literature on both types of variables. The idea is to expose how various studies have condemned the exclusive use of quantitative variables (i.e. financial ratios) established the necessity of the inclusion of qualitative variables in building CF-IPMs. After establishing this necessity, it is believed that the fairest investigative way forward is to review as many as possible studies that have developed CF-IPMs to see how many included qualitative variables as a result; this was done using a systematic review.

To achieve part of the first objective listed in chapter one which is 'to identify and collect data on qualitative variables that contribute to solvency/insolvency of construction firms through literature review and fieldwork', a systematic review of studies on the failure of construction firms was also done. This review can help reveal potential qualitative variables that can be used to build a robust CF-IPM in the research.

The next section (i.e. section 3.2) is a state of the art literature review on the necessity of inclusion of qualitative variables in CF-IPMs. Section 3.3 is a systematic review of studies on construction firms insolvency prediction model where the studies that have used qualitative variables in any form were identified. Section 3.4 is a systematic review of studies on failure of construction firms to unmask potential qualitative variables needed to build a high-performance CF-IPM in the research. Section 3.5 present a summary of discoveries made from the rigorous reviews in the chapter

3.2 Types of variables used in construction firms insolvency prediction models

As much as owners do not like to hear it, the high prospect of construction firm insolvency, in any case, is a real one. The negative impact of such insolvencies on the economy and society, in general, has led to the development of many insolvency prediction models. However, the effectiveness of an insolvency prediction model is dependent on, amongst other factors, the variables that are chosen to develop it.

The exclusive use of financial ratios (quantitative variables) as variables is common with virtually all insolvency prediction models (IPMs), including those built for construction firms (e,g. Mason and Harris 1979; Langford *et al.* 1993; Abidali and Harris 1995; Singh and Tiong 2006; Thomas Ng *et al.* 2011; Horta and Camanho 2013 among others). The method (i.e. the exclusive use of financial ratios) is a case of ‘follow the crowd’ approach for most IPMs as this has been the prevailing method since the days of the pioneering IPM studies (i.e. Beaver 1966; Altman 1968). The method is also attractive because of the ease with which data, which is financial ratios, can be collected. Submission of annual financial statements, which contains elements for calculating financial ratios, is a legal requirement for registered (construction) firms; these submissions are normally readily available in third party databases. In the United Kingdom, for example, the financial data of all firms are available with ‘Company House’ for a token, an executive agency and trading fund of Her Majesty's Government, which serves as the United Kingdom's registrar of companies. The information is also available on databases such as FAME (Forecasting Analysis and Modelling Environment) Bureau Van Dijk UK.

Qualitative variables, on the other hand, are less popular because they are not readily available anywhere and information on them might require the rigorous task of interviewing surveying managers of many sample firms. This can be a Herculean task especially because it involves finding representatives of failed construction firms who can be very difficult to track down and might be unready to discuss their firm's failure if it causes them pain. However, they (i.e. qualitative variables) remain extremely important and must be used if a robust prediction model is to be developed, especially for construction firms. This has been reiterated in many studies.

Right from the very early stages of IPM research, Argenti (1976) who is a well-known professor of corporate failure proved the multiplicity of causes of failure by clearly highlighting how major books and studies on firm failure highlighted entirely different many reasons (as many as 10 in many cases) for failure of firms. He (Argenti, 1976) consequently argued that financial ratios are incapable of exclusively predicting failure as there are multiple causes. Argenti (1976) clearly established that the failure of firms, big or small, is not a sudden process, clarifying that micro or small and medium-sized enterprises (SME) can take years to fail while large firms can take decades. He noted the non-financial factors (i.e. qualitative variables) like defects in the top management team could be used to detect potential failure earlier. The defects highlighted included autocratic chief executive, many inactive board members, chief executive acting as chairman, unbalanced skills and knowledge of the board members, among others. Fraud and bad luck, which is agreed to cause 1% of firms' failure, were also identified as non-financial factors

Argenti (1976) explained that it is 'creative accounting' that causes a firm that is nearing failure to look stable. He (Argenti, 1976) gave an example of creative accounting where he explained about Rolls Royce that "because the company capitalized the annual R and D costs of the RB211 the company could demonstrate a fairly healthy profit and hence justify the continued dividend payments. In fact, it was making a loss.But no one wanted to know" (p. 13). This creative accounting results in financial ratios that are not true representations of a firm's situation by depicting a very stable firm when it is failing. This further calls into question the exclusive use of financial ratios for IPMs.

Kangari (1988), an active researcher and then associate professor in construction engineering and management, looked into the failure of construction firms in particular. He

(Kangari, 1988) was able to highlight very early (in 1988) that what was needed to prevent or reverse a construction firm's impending failure was to establish common features of failing construction firms find a way of synthesising them. The emphasis in Kangari's work was to understand "the mechanism behind the financial failure of construction companies" (Kangari 1988, p.173) rather using the financial information to understand failure. In essence, it is non-financial activities/events (i.e. qualitative variables) that actually dictate what the financial situation (or financial ratios) of a firm will look like. It is thus important to use qualitative variables if early prediction of failure that will allow some time for remediation is to be achieved

Kale and Arditi (1999) lamented the fact that construction management research on construction firms' failure is mainly approached from the perspective of finance (quantitative variables) while no one paid attention to the organisational theory perspective (qualitative variables). They noted Hall (1994) as the only study which considered "the process which leads to construction business failures" and "explored the factors associated with the failure of construction companies" (Kale and Arditi 1999; p. 494). [Hall's (1994) work will be discussed in the next section where a systemic review of CI IPM studies is done]. Kale and Arditi (1999) highlight "the fragmented nature of the industry structure, the fragmented nature of the organization of the construction process, easy entry to the construction business, post-demand production, the one-off nature of projects, the high uncertainty and risk involved, the high capital intensiveness of the constructed facilities and the temporary nature and duration of exchange relationships" (p. 496) as characteristics that make the CI unique and hence the need to look beyond accounting variables to understand failure of its (i.e. the CI) firms

Arditi *et al.* (2000) explained that studies that developed IPMs in the three decades to the year 2000, unfortunately, did so exclusively with financial ratio and expressed the same regret as Kale and Arditi (1999) about the scarcity of use of organisation theory, which requires qualitative variables. Arditi *et al.* (2000) condemned the exclusive use of financial ratios for CI IPMs by citing Argenti (1976)'s argument that financial ratios cannot detect the causes of failure and might be unreliable due to 'creative accounting' practices though they can detect some of the associated warning signs. Arditi *et al.* (2000) concluded that incorporating "data based on both organisational and managerial foundations rather than on financial ratios is still open to researchers in the construction industry" (p. 120). He (Arditi, Koksal and Kale, 2000) went ahead to “explore the factors”, but his work did “not include the development of an empirical model for predicting business failure” (p. 120).

This incisive review reveals how reputable studies have established that prediction of failure of construction firms cannot be well made with the exclusive use of financial information. The calls to use qualitative variables have been loud and clear. In order to establish CI IPM studies that have in any way attempted to use non-financial qualitative variables, a systematic review of CI IPM studies.

3.3 Systematic review of construction firms insolvency prediction model studies with focus on variable types

“A systematic review is a summary of the research literature that is focused on a single question. It is conducted in a manner that tries to identify, select, appraise and synthesise all high-quality research evidence relevant to that question.” (Bettany-Saltikov 2012: p.5). The systematic literature review method obligates a broad search of the literature (Smith *et al.*, 2011) with an unambiguous expression of exclusion and inclusion criteria (Nicolás and Toval 2009). Systematic review is renowned for yielding valid and repeatable/reliable results because it reduces bias to a minimum hence its high recognition and frequent use in the all-important medical research world (Tranfield, Denyer and Smart, 2003; Schlosser, 2007) and its embracement in other research areas like IPM (Appiah, Chizema and Arthur, 2015). The general review of various existing knowledge and synthesising them is also a recognised method which contributes immensely to the progression and expansion of knowledge (Aveyard, 2014; Fink, 2010). This is the reason it has been widely employed as a methodology in various research areas including insolvency prediction (Balcaen and Ooghe 2006; Adnan Aziz and Dar 2006) and construction business failures (Edum-Fotwe, Price and Thorpe, 1996; Mahamid, 2012).

Since results from peer reviewed journals are generally considered to be of high quality and validity (Schlosser, 2007), this systematic review employs only peer-reviewed journals. This will ensure a high validity of the review results. The databases searched for this review include Google Scholar (GS); Wiley Interscience (WI); Science Direct (SD); Web of

Science UK (WoS); and Business Source Complete (BSC). This is done in tandem with the latest published systematic review article on IPM (i.e. Appiah *et al.* 2015). Observations revealed that GS, WoS and BSC contained all the journal articles provided in Wiley and Science Direct since the later are publishers while the former are general databases. To further broaden the search, the Engineering Village (EV) database was added to the GS, WoS and BSC databases to perform the final search.

Pilot searches revealed that studies use bankruptcy, insolvency and financial distress interchangeably to depict failure of firms. A search structure which included all these words was subsequently designed with the following defined string ("Forecasting" OR "Prediction" OR "Predicting") AND ("Bankruptcy" OR "Insolvency" OR "Distress" OR "Default" OR "Failure") AND ("Construction" OR "Contractor"). A process flow of the systematic review methodology is presented in Figure 3.1.

To avoid database bias, ensure high repeatability and consistency of the research, and consequently high reliability and quality, all the relevant studies that emerged from searching the databases were employed in the review (Schlosser, 2007). Since the databases host studies from around the globe, geographic bias was readily averted. Considering that the first set of IPM studies emerged in the 1960s (Beaver, 1966; Altman, 1968), a period of 1960-2015 (the year this review was done) was used for the search.

One of the inclusion criteria was for the IPM study to focus solely, or mainly, on the CI. Another is that the study must employ quantitative factors (i.e. financial ratios as variables). The titles and abstracts of the studies that the search returned were typically adequate to decide the ones qualified for use in the research. Where otherwise, articles' introduction and conclusion were read to determine their suitability. The extent of reading was dependent on the information got from initial readings. In exceptional cases, the full-length article was read. In the end, GS produced 31 results, EV (14), BSC (11) and WoS (7). Most of the articles returned in searching EV, BSC and WoS were present in the GS search results. In fact, all EV results were present in the GS result, while BSC and WoS were only able to produce four and one unique articles respectively

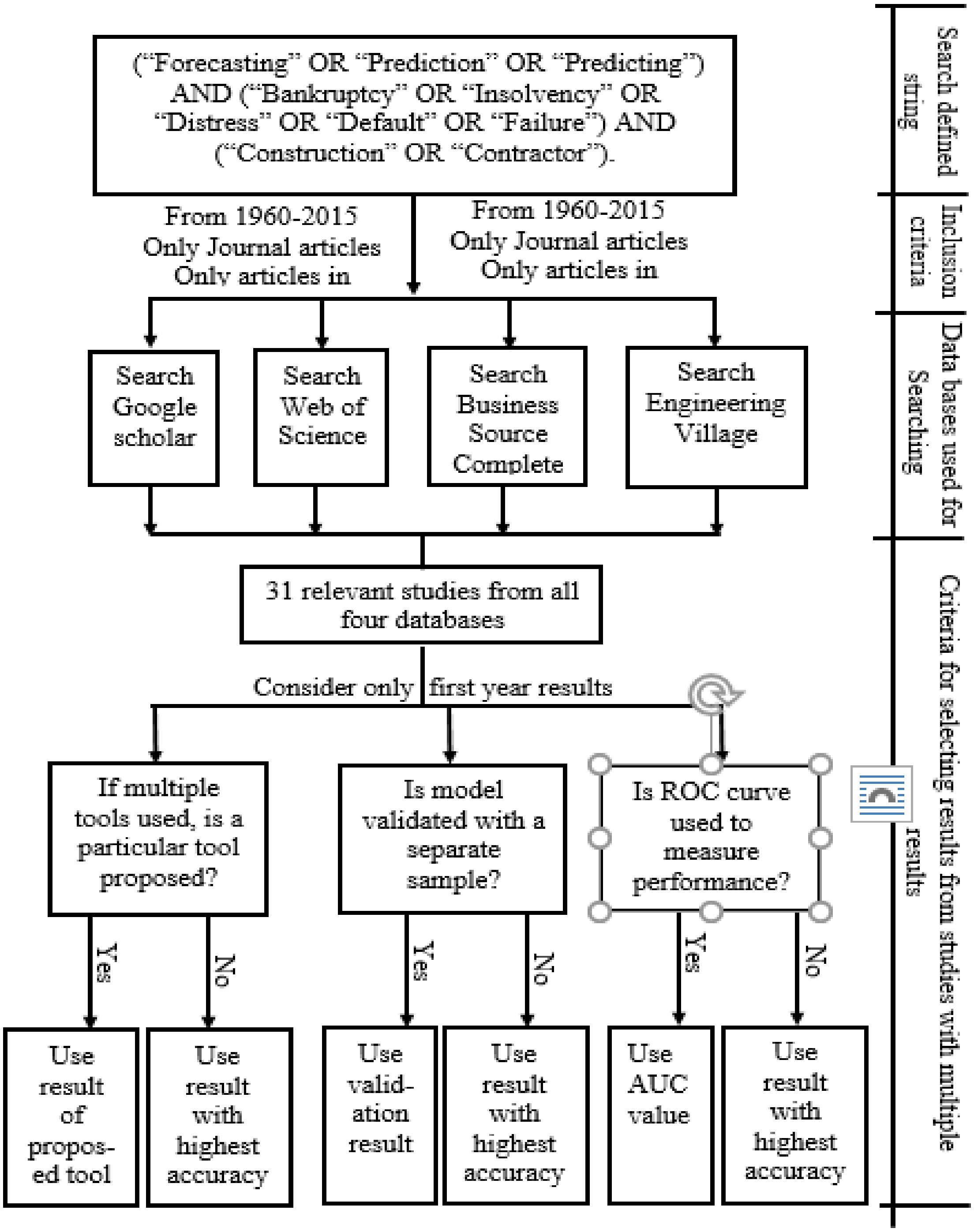

("Forecasting" OR "Prediction" OR "Predicting") AND ("Bankruptcy" OR "Insolvency" OR "Distress" OR "Default" OR "Failure") AND ("Construction" OR "Contractor").
From 1960-2015
Only Journal articles
Only articles in
From 1960-2015
Only Journal articles
Only articles in
Search Google scholar
Search Web of Science
Search Business Source Complete
Search Engineering Village
31 relevant studies from all four databases
Consider only first year results
If multiple tools used, is a particular tool proposed?
Is model validated with a separate sample?
Is ROC curve used to measure performance?
Yes
No
Yes
No
Yes
No
Use result of propos-ed tool
Use result with highest accuracy
Use valid-ation result
Use result with highest accuracy
Use AUC value
Use result with highest accuracy
Search defined string
Inclusion criteria
Data bases used for Searching
Criteria for selecting results from studies with multiple results

Figure 3.1: A process flow of the systematic review methodology for CI IPM studies variables

ROC: Receiver Operating Characteristic

AUC: area under the curve

The exclusion criteria included, among others, articles that were not written in the English language. Although language constraint is not favoured in systematic review, it is unavoidable and thus acceptable when there is a lack of funds to pay for interpretation services (Smith *et al.*, 2011). An example of study excluded based on language is Wedzki (2005) which is written in Polish. Review studies were not considered as they did not develop any model and hence cannot be said to have used any particular type of variables. Unsuitable articles with titles like ‘default prediction for surety bonding’ (e.g. Awad and Fayek 2013) and ‘contractor default prediction before contract award’ which fixate on a contractor’s capability to successfully execute a specific kind of project (e.g. Russell and Jaselskis 1992) were taken out. After this step, only 28 studies were left. Note that ‘contractor default prediction before contract award’ articles that fixated on insolvency probability as the main/only judging criteria were not excluded as the studies effectively built a form of CF-IPM.

In the final 31 articles reviewed in this section, where multiple accuracy results are presented for multiple CF-IPMs, only the accuracy result of the proposed tool in the article is presented in the research. Where no particular tool is proposed, the highest accuracy result is presented here. Where the results for training and validation samples are given, the validation result is used here. Otherwise, the training result is adopted. Where error types are calculated independent of accuracy values, and the Receiver Operating Characteristic (ROC) curve is used to determine performance, the area under the curve (AUC) value in percentage is taken as the accuracy result. Where accuracy results of multiple years are given, the result of the first year is adopted to allow fair comparison since the first year result is the most commonly presented result in IPM studies. As required for systematic review, a meta-analysis based on variables used was done with data synthesised through the use of ‘Summary of Findings’ tables (Higgins and Green 2008; Smith *et al.* 2011) in Table 3.1.

Looking at the table, no special analysis or statistics is needed to find out that only four (Hall 1994; Abidali and Harris 1995; Koksal and Arditi 2004; Horta and Camanho 2013) out of 31 studies used some form of non-financial qualitative variables in their studies, and each study had issues.

Table 3.1: Summary of findings table for quantitative factors

S/N	Author (Year)	Financial (quantitative) variables category used	Non-financial (qualitative) variables category used	Accuracy of CF-IPM (%)
1.	Fadel (1977)	Profitability		-
2.	Mason and Harris (1979)	Profitability, working capital position (liquidity), leverage, quick assets position, trend.		87
3.	Kangari and Farid (1992)	Profitability, Efficiency, Liquidity		-
4.	Langford *et al.* (1993)	Short term solvency, solvency, liquidity, profitability,		63.33
5.	Hall (1994)		Firm characteristic, Management/ Owner Characteristics, Skills of the Workforce, Management decision making, Motivation	
6.	Abidali and Harris (1995)	Profitability, Leverage, Activity/net asset, turnover, Liquidity, Trend measurement	Management/ Owner Characteristics, Skill of workforce, Management decision making, Internal Strategic	70.3
7.	Russell and Zhai (1996)	Trend, future position, volatility		78.3
8.	Koksal and Arditi (2004)		Management/ Owner Characteristics	
9.	Singh and Tiong (2006)	Short-term liquidity, cash position (cash flow), long-term solvency, profitability, managerial performance		

S/N	Author (Year)	Financial (quantitative) variables category used	Non-financial (qualitative) variables category used	Accuracy of CF-IPM (%)
10.	Chen (2009)	Short-term liquidity and efficiency, Capital structure and Solvency, Profitability and market prospect, Economy		86.13
11.	Huang (2009)	Leverage, solvency, liquidity		88.5
12.	Sueyoshi and Goto (2009)	profitability, leverage, growth, size and risk		93.9
13.	Stroe and Bărbuță-Mișu (2010)	Profitability, solvency, liquidity, rate of financial expenses, rate of personnel costs		77.8
14.	De Andrés *et al.* (2011)	Liquidity, profitability, leverage, management efficiency		88.72
15.	Ng *et al.* (2011)	Operation, profitability, solvency and cash flow		96.9
16.	Tserng *et al.* (2011)	Stock market information		90%
17.	Tserng *et al.* (2012)	Management efficiency, solvency, leverage, activity ratio (management)		84.5
18.	Tserng, Lin, *et al.* (2011)	Management efficiency, liquidity, asset utilization.		80.31
19.	Chen (2012)	Profitability, management efficiency, growth, leverage, activity ratio, asset utilisation (management), liquidity		85.1
20.	Sánchez-Lasheras *et al.* (2012)	Liquidity, profitability, asset utilization (management), leverage, management efficiency		89.58
21.	Tsai *et al.* (2012)	Profitability, Activity, Leverage and Liquidity		87.32
22.	Horta and Camanho (2013)	profitability, liquidity, leverage, activity	Company characteristics	97.6

S/N	Author (Year)	Financial (quantitative) variables category used	Non-financial (qualitative) variables category used	Accuracy of CF-IPM (%)
23.	Makeeva and Neretina (2013a)	Liquidity, Turnover, Profitability, Financial Solidity		86.44
24.	Makeeva and Neretina (2013b)	profitability, liquidity, and turnover measures, size and interest coverage coefficients		86.44
25.	Sun *et al.* (2013)	Solvency, Profitability, Activity, Per-share ratio, Structural ratio, Growth ratio		93.07
26.	Cheng *et al.* (2014)	Liquidity, Activity, Profitability		92.13
27.	Heo and Yang (2014)	Working capital utilization (management), liquidity, asset utilization (management), asset structure		78.5
28.	Muscettola (2014)	Management efficiency, profitability, solvency, interest coverage ratio		80.94
29.	Tserng *et al.* (2014)	liquidity, profitability, leverage, activity and market factor		79.18
30.	Cheng and Hoang (2015)	liquidity, leverage, activity, and profitability		96.0
31.	Tserng *et al.* (2015)	Profitability, leverage, leverage group.		84.8

Management efficiency include asset utilisation, activity ratio, working capital utilisation

Solvency ratio is the same as leverage

Growth ratios are a form of trend ratio

Analysing the studies in chronological order, Hall (1994) was the most comprehensive user of qualitative variables. Hall interviewed 28 and 30 owners of failed and existing small construction firms (i.e. with less than 100 employees) respectively, with the analysis of interviews leading to 93 potential variables. Only firms in the North-west of England were involved. An attempt to group the variables into principal components for a Logit model using Equamax rotation failed as it yielded only 14 components with a maximum number of two variables in each component. A stepwise Logit model was subsequently used "separately on each half of the variable set (odd and even numbered variables respectively)" (p.747) leading to a Logit model with six variables.

Although Hall's (1994) work was the most distinctive and arguably the best in that it was able to focus on small construction firms and used the required qualitative variables for this purpose, it still had some major flaws. About the research, the first and most important was its total avoidance of financial ratios. Despite their flaws (i.e. financial ratios), they have been proven by many studies (e.g. Altman 1968b; Mason and Harris 1979; Kangari and Farid 1992; Russell and Zhai 1996; among others) to make some vital contributions to failure prediction hence combining them with qualitative variables is the best option. In fact, it is the exclusive use of financial ratios that has been condemned (Argenti, 1976) and not its inclusion in the first place

Also, Hall (1994) used interviews of 58 construction firm owners (28 and 30 existing and failed respectively) to identify 93 variables but did not explain how the number value of the variables for each firm was gotten before being used in a Logit model. Were the respondents told to rate their firms in percentage or on a given scale (e.g. one to seven) for each variable? Or did Hall (1994) just assign each variable with a value for each firm based on how the respondents responded to his questions? This remains a myth.

Further question marks over Hall's (1994) work include the claim of 95% accuracy achieved by the model as this was achieved with the data used to build the model rather than a separate test data. In essence, the model was neither validated nor tested hence any reported accuracy is unrealistic (see chapter five on review of methods). Finally, no justification for concentration on only firms from the North-west of England was given.

The next study that employed qualitative variables in Table 3.1 is that of Abidali and Harris (1995). Although Abidali and Harris (1995) used both quantitative and qualitative variables as indicated in Table 3.1, they were used for separate models (Z-score and A-score models respectively) rather than combined. Abidali and Harris (1995) carried out a case study of failed firms to identify qualitative variables that can be used to predict failure and came up with 13 variables that are quite similar to ‘the management defects’ identified by Argenti (1976). A survey of a separate 28 firms was done on the 13 variables, and each variable was given a weight based on the percentage that ticked it as a contributing factor to failure of construction firms. Then the variables “were identified from the survey sample for 'solvent 7' and 'at risk 7' groups based on their Z scores” (p. 194). This identification, together with the assigned weights based on the percentage responses, were used to create the A-model i.e. the model based on qualitative variables.

The flaws in Abidali and Harris' (1995) attempt to use qualitative variables are similar to those of Hall's (1994) if not more. First Abidali and Harris (1995) did not combine the variables. Secondly, they did not explain how the number value of the variables for each firm was gotten. Instead, they simply said the variables “were identified from the survey sample for 'solvent 7' and 'at risk 7' groups based on their Z scores” (p. 194). What makes the model worse, in this case, is that the weights assigned to each variable were assigned using an unestablished statistical process of response percentage. Finally, the model was equally neither tested nor validated.

Koksal and Arditi (2004), in a similar way to Hall (1994) used only qualitative variables but did not focus on small firms. Their (i.e. Koksal and Arditi 2004) approach was much more valid regarding methodology as respondents were required to “rate each potential cause of decline… relative to the existing conditions in their company using a five-point scale where “1 = extremely weak,” “2 = weak,” 3 = fairly strong,” “4 = strong,” and “5 = extremely strong.”” (p.803). In essence, Koksal and Arditi (2004) did not arbitrarily or use personal discretion to assign number values to variables. However, a major flaw is that Koksal and Arditi (2004) failed to include and construction firms/industry specific variables that can really indicate the potential future of construction firms in their 21 variables. Instead, the focus was on general management variables based on ‘organisation theory’. This makes their model somewhat generic as it has no specificity to the construction firms. Also, even though Koksal and Arditi (2004) clearly understand that “ideally, one should test a model by using cases that are independent of the cases used in the development of the model… this was not possible in the research” (p. 806). In essence, the model was more or less neither validated nor tested.

Horta and Camanho (2013) made the first attempt to combine quantitative and qualitative variables in a CF-IPM. The method is however flawed in that the qualitative variables were chosen arbitrarily and there were only three of them: main company activity, company size and headquarter geographic location. The arbitrariness means there is zero proof of the effectiveness of the variables. Looking at the variables chosen, it is clear that the authors (i.e. Horta and Camanho 2013) simply decided to use variables that are very easy to get from free sources such as financial statements, company websites, among others. The qualitative variables are at best, very poor when compared with the three initially discussed studies (i.e. Hall 1994; Abidali and Harris 1995; Koksal and Arditi 2004).

Of the four studies, the only study (i.e. Horta and Camanho 2013) that used an Artificial Intelligence (AI) tools used only one tool (i.e. support vector machine) and thus could not compare models between different powerful AI tools. Other studies simply used the popular logit regression, also known as logistic regression. So overall, none of the studies that used qualitative variables in their models used and compare models from various powerful AI tools.

The process in this research will address these identified flaws by combining quantitative and qualitative variables in various powerful AI tools thus allowing for comparison of models to allow selection of the best model. The qualitative variables will be generated based on literature review and interviews rather than arbitrarily while the numeric value of each variable in relation to each sample firm will be gotten through questionnaire survey responses of owners and directors of failed and existing construction firms. To identify potential qualitative variables from literature, a systematic review of studies that have dealt with the failure of construction firms using qualitative factors/variables is carried out in the next section. This will include the four studies discussed above (i.e. Hall 1994; Abidali and Harris 1995; Koksal and Arditi 2004; Horta and Camanho 2013)

3.4 Systematic review of studies on failure of construction firms

The systematic review of the studies that have dealt with failure of construction firms using qualitative factors/variables was done in a similar way to that of the CF-IPM studies review except for a few differences which are explained here. A search structure with the following defined string was designed: (“Business” OR “Firm” OR “Company”) AND (“Bankruptcy”

OR “Insolvency” OR “Distress” OR “Default” OR “Failure”) AND (“Construction” OR “Contractor”). A process flow of the systematic review methodology for qualitative factors is presented in Figure 3.2.

Only eight suitable articles were found, in addition to the previously identified four (i.e. Hall 1994; Abidali and Harris 1995; Koksal and Arditi 2004; Horta and Camanho 2013), after a strenuous inspection of more than 500 articles. The result was improved by checking review articles and checking through their citations/references. Three more studies (Jannadi 1997; Robinson and Maguire 2001; Arslan *et al.* 2006) were added using this method. With no resulting article identifying the role of environmental, social and governance (ESG) in failure of construction firms, the search words ‘sustainability practices and failure of construction companies’ were used on Google, and the first suitable article (i.e. Siew *et al.* 2013) was selected. As a result, a total of 15 primary studies was reviewed altogether. As required for systematic review, a meta-analysis was done with data synthesised through the use of ‘Summary of Findings’ tables (Higgins and Green 2008; Smith *et al.* 2011) in Table 3.2.

The result of the systematic review, as presented in Table 3.2, represent the potential qualitative variables that can be used to build CF-IPMS. The factors identified in the systematic review coupled with the responses from the interviews will be carefully analysed and used to create a questionnaire of qualitative variables. Please see chapter six for further details.

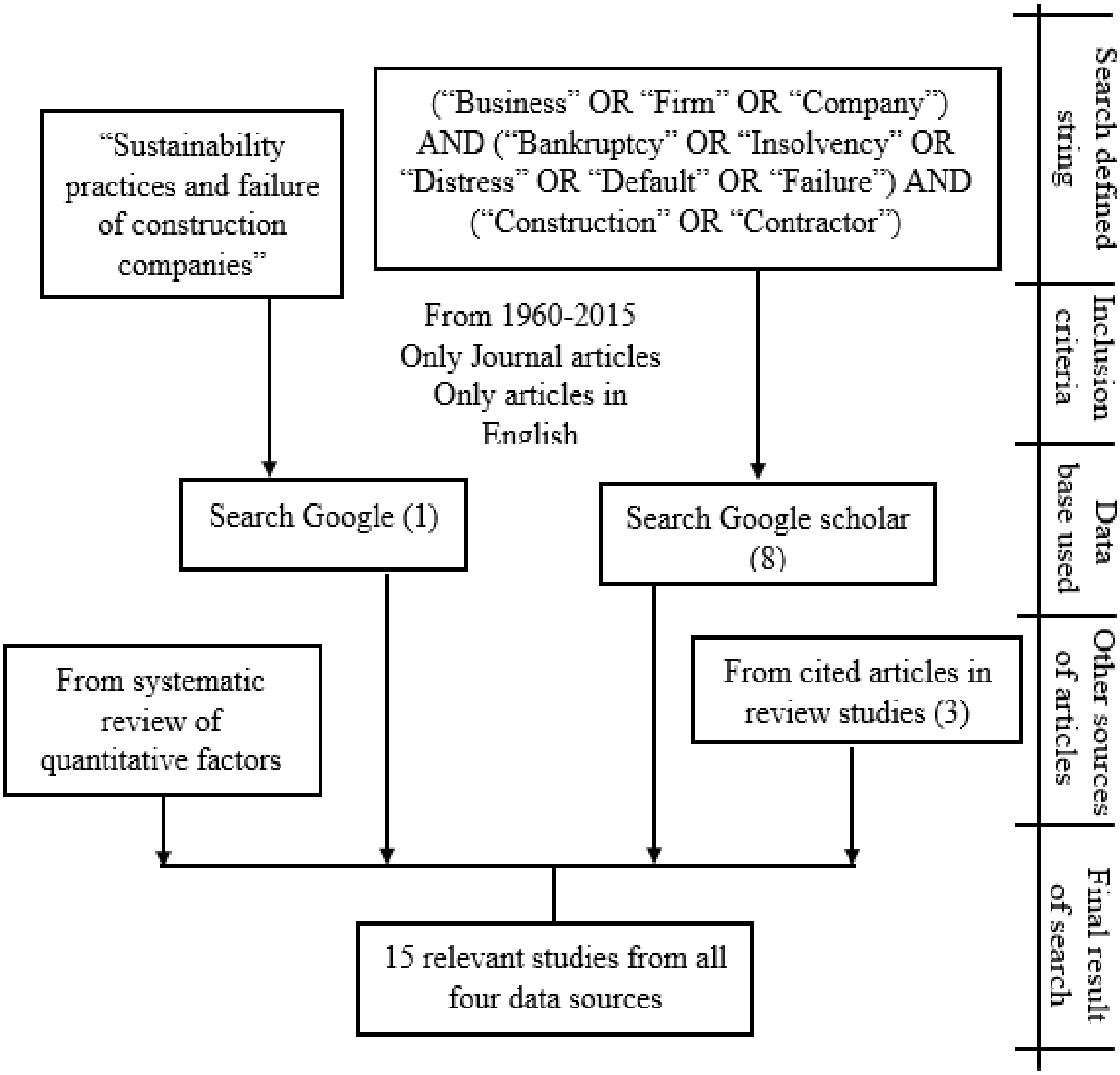

"Sustainability practices and failure of construction companies"
("Business" OR "Firm" OR "Company") AND ("Bankruptcy" OR "Insolvency" OR "Distress" OR "Default" OR "Failure") AND ("Construction" OR "Contractor")
From 1960-2015
Only Journal articles
Only articles in English
Search Google (1)
Search Google scholar (8)
From systematic review of quantitative factors
From cited articles in review studies (3)
15 relevant studies from all four data sources
Search defined string
Inclusion criteria
Data base used
Other sources of articles
Final result of search

Figure 3.2: A process flow of the systematic review methodology for qualitative factors

Table 3.2: Summary of findings table for qualitative factors

S/N	Author Year	Qualitative variables employed	Factors (Variables category)
1.	Kangari (1988)	Rate of construction activity, Inflation, Influx of new firms	Macroeconomic (including industry), Economic
2.	Hall (1994)	Age, Education of owner, Skills of the Workforce, Use of Information (especially accounting info in decision making), Management of Cashflow, Motivation, Ploughed-back profit	Firm characteristic, Management/ Owner Characteristics, Skills of the Workforce, Management decision making, Motivation
3.	Abidali and Harris (1995)	autocratic chief executive, the same person as both chief executive and chairman, the company board, lack of engineering skills, lack of a strong financial director, defective managerial skills, incomplete accountancy system, defective bidding system, poor marketing skills, over-trading, losses in projects, Acquisition of a potentially failing firm	Management/ Owner Characteristics, Skill of workforce, Management decision making, Internal Strategic
4.	(Russell and Zhai 1996)	Value of new construction put in place, value of construction contracts	Firm characteristics (size)
5.	(Jannadi, 1997)	Difficulty in acquiring work, bad judgment, lack of experience in the firm's line of work, difficulty with cash flow, lack of managerial experience, and low-profit margins.	Management decision making, Skill of workforce

S/N	Author Year	Qualitative variables employed	Factors (Variables category)
6.	(Kale and Arditi 1999)	Age, organisational learning, size, the turbulence of the construction industry, gaining legitimacy (i.e. company's image)	Firm characteristics, Macroeconomic (including industry)
7.	(Arditi, Koksal and Kale, 2000)	Insufficient profits (heavy operating expenses, insufficient capital and burdensome institutional debt. All	Budgetary/finance
		Industry weakness	Macroeconomic (including industry)
		Lack of business knowledge	Management/ Owner Characteristics
8.	(Robinson and Maguire 2001)	growing too fast, obtaining work in a new geographic region, dramatic increase in single job size, obtaining new types of work, high employee turnover, inadequate capitalization, poor estimating and job costing, poor accounting system, poor cash flow, and buying useless stuff.	External strategic, Internal Strategic, Management decision making
9.	**Enshassi *et al.* (2006)**	Lack of experience in the line of work	Skill of workforce/
10.	(Arslan *et al.*, 2006)	difficulties with cash flow and poor relationship with the client drove the contractor failure, poor bid proposal that drives down profit	Management decision making, Internal Strategic

S/N	Author Year	Qualitative variables employed	Factors (Variables category)
11.	(Koksal and Arditi 2004)	competitive strategy, defining competitive advantage, adaptation to advanced managerial practices, and adaptation to advanced construction technologies	Internal Strategic
		absence of standardisation, defining the scope of the company, diversification of the production markets, and absence of specialisation	Internal Strategic
		Level of business knowledge, level of work experience and level of managerial experience	Management/ Owner Characteristics
12.	**Dikmen *et al.* (2010).**	Management incompetence, Poor value chain analysis at the corporate level, Poor strategic planning, Poor environmental scanning, Poor financial management Poor leadership, Poor human resource management, Poor communication, Poor planning and scheduling, Poor monitoring and control, Poor organisation of resources, Poor selection and management of supply chain, Poor quality management and control, Poor project risk management, Poor change order and claim management	Management decision making, Management/ Owner Characteristics, Internal Strategic
		Lack of organisational knowledge, Poor technical and technological capacity, Poor relations with clients/government, Poor company image	Internal Strategic, Firm characteristics
		Unsuccessful restructuring/reorganisation, saving non-value adding activities, Poor investment decisions, Wrong level of diversification, Wrong project selection, Poor project cost estimation, Excessive expansion	External strategic, Management decision making, Management/ Owner Characteristics
		Difficulty in collecting money from the client, Unexpected change within the workforce, Sudden death of the company leader, Economic fluctuations, Shrinkage in construction demand, Change in politics	Macroeconomic (including industry), Management/ Owner Characteristics

S/N	Author Year	Qualitative variables employed	Factors (Variables category)
13.	**Holt (2013)**	Rapid growth	Rapid growth
		Inadequate capitalisation, Poor management, Requirement for cash flow, Poor systems, poor cost control	Management/ Owner Characteristics, Management decision-making
		Economic	External strategic
		Economic/market conditions	Macroeconomic (including industry)
		Strategic errors contributing to failure, Catastrophe builds, management inaction, Single minded company attitude, Self before the company, Over-optimism/failure to perform, Financial failure from managerial performance	Internal Strategic, Management/ Owner Characteristics, Management decision-making
		Liability of newness, Poor relationships with SMEs and customers, Poor intelligence or decision data, Inaccurate or improper financial reporting, Poor trained management	Firm characteristics, Management/ Owner Characteristics, Management decision-making
14.	(Horta and Camanho 2013)	company main activity, company size and headquarter geographic location	Firm characteristics
15.		Direct emissions from facility or process, including those occurring in equity stakes, Indirect emissions associated with purchased electricity, Supply-chain carbon emissions, Opportunities to pass carbon costs on to customers, Opportunities to reduce carbon emissions and energy use, among others	Sustainability
		Monetary values of fines and number of non-monetary sanctions for noncompliance with environmental laws and regulations, Environmental provisions as reported on the balance	Sustainability

S/N	Author Year	Qualitative variables employed	Factors (Variables category)
	(Siew, Balatbat and Carmichael, 2013)	sheet, Number and severity of transgressions of environmental license conditions, among others	
		Type of waste produced by product and volume, Targets for the reduction of waste, % of waste re-used in the manufacturing process, Water consumed (by quality/source) and targets for reduction, % water recycled compared with base year, among others	Sustainability
		Hazardous waste emissions and reduction, NOx, SOx and particulate emissions, Emissions of ozone-depleting substances by weight, Total water discharge by quality and destination, Details of toxic materials used in the manufacturing process, among others	Sustainability
		Training courses offered or held, Audits actually conducted by independent parties; Monitoring conducted/initiatives; Incidents analysed breakdown, Number of near misses reported, % of hazards rectified	health and safety
		Board oversight of HCM, Integration of HCM and people risks into risk management processes, Executive remuneration linked to achievement of HCM objectives, Employee diversity/anti-discrimination policies, Processes to monitor and address discrimination, among others	Management decision-making
		Corporate codes of conduct, the extent of their application and associated training, Responsibility within the organisation for the code of conduct, Linkages between remuneration policies and code of conduct, among others	Firm characteristics
		Basis for identifying the key stakeholders with which to engage, Frequency of key stakeholder engagement, Engagement mechanisms, Main issues arising from stakeholder engagement, Steps taken to respond to stakeholder feedback	Management decision-making
		Risk management policies and implementation, The boards' assessment of related party issues, Director selection and board succession planning process, among others	Management decision-making

3.5 Chapter summary

This chapter presented a literature review of the variables used in building CF-IPMs from the perspective of quantitative (i.e. financial ratios) and qualitative variables. Without building any form of CF-IPMs, Argenti (1976), Kangari (1988), Kale and Arditi (1999), and Arditi *et al.* (2000), among others, strongly established the need to include qualitative variables in developing CF-IPMS if they are to be very valid

A systematic review of the studies that have built CF-IPMs revealed that only four (Hall 1994; Abidali and Harris 1995; Koksal and Arditi 2004; Horta and Camanho 2013) out of 31 reviewed studies, used some form of non-financial qualitative variables, and each study had issues. Issues identified include non-combinatory use of quantitative and qualitative variables (i.e. exclusive use of qualitative variables) in three studies four (Hall 1994; Abidali and Harris 1995; Koksal and Arditi 2004); unexplained or arbitrary assignment of values to variables (Hall 1994; Abidali and Harris 1995); non-use of advanced tools i.e. artificial intelligence tools like artificial neural network, support vector machine, (Hall 1994; Abidali and Harris 1995; Koksal and Arditi 2004), among others.

The method of the only study that attempted to combine both variables (i.e. Horta and Camanho 2013) was flawed in that the qualitative variables were chosen arbitrarily and there were only three of them: main company activity, company size and headquarter geographic location. Arbitrariness translates to zero proof of the effectiveness of the variables. The selection was clearly based on variables that are very easy to get from free sources such as financial statements, company websites, among others. The qualitative variables are at best, very poor when compared with the three initially discussed studies (i.e. Hall 1994; Abidali and Harris 1995; Koksal and Arditi 2004). Horta and Camanho (2013) use of only one AI tool also meant they could not compare models between different powerful AI tools.

These flaws will be properly addressed in the research. The last section of this chapter was used to systematically review studies on failure of construction firms thereby revealing many potential qualitative variables that can be used to the build the CF-IPM of the research.

In chapter four, an explanation of the concept of Big Data Analytics (BDA) and Big Data Engineering, as well as the justification for the use of BDA in the research were given.

CHAPTER FOUR

4.0 BIG DATA FOR CONSTRUCTION FIRMS INSOLVENCY PREDICTION

4.1 Chapter introduction

Big Data remain a ‘catch phrase’ in the present world and retains some level of contemporariness despite having been around for over five years now. It is even a lot newer to the construction industry that has not yet taken full advantage of its superior analytics capabilities. Big Data Analytics can be useful to the construction industry in many ways, one of which is in the area of construction firms insolvency prediction models (CF-IPM) research. The CF-IPM and the general insolvency prediction models (IPM) research areas have long struggled with the ability to combine the use of a large amount of data and well-tuned sophisticated artificial intelligence tools. An attempt to do this by Du Jardin (2010) resulted in onerous computation that took a duration of five days with 30 PCs running Windows. Such a difficulty has no place in the contemporary world where Big Data Analytics can be used to do the same purported ‘onerous computation’ in minutes or seconds.

This chapter is an explanation of Big Data Engineering and Analytics and how they relate to construction firms insolvency prediction models (CF-IPM). It explains what makes a data qualify for Big Data and describes how the limitation of the most common Big Data processing model makes it unfit for CF-IPMS. The available options to solving problems as well are discussed in the context of CF-IPM. The approach to be used in the research and the associated reason, and why the data available qualifies as Big Data in the first place, are all explained.

Section 4.2 introduces Big Data from its very origin and its two aspects: engineering and analytics. Section 4.3 is about Big Data Engineering. Subsection 4.3.1 is a description of Big Data’s most common programming model (or executioner) which is Hadoop MapReduce. The subsection is also used to highlight classification through iteration, which is the basic back-end processor of any CF-IPM, as a major problem of Hadoop MapReduce. Subsection 4.3.2 is used to describe the available initiatives that can help solve the classification through iteration problem, with 4.3.2.1 and 4.3.2.2 focusing on MapReduce and Spark based initiatives respectively. Subsection 4.3.3 is used to present and justify the choice initiative that will be used to develop a Big Data CF-IPM in the proposed solution. It is also used to provide two frameworks: one for selecting the right initiative for other users and the other for developing CF-IPM with Big Data Analytics. Section 4.4 is about Big Data Analytics. Section 4.5 is a justification of why the available data is suitable for Big Data Analytics. Section 4.6 is a summary of this chapter

4.2 Introduction to big data

The combo of words 'Big Data' was coined by John Mashey who first used it in his Silicon Graphics (SGI) slide titled "Big Data and the Next Wave of InfraStress" (Diebold, 2012b). Though Big Data definition is complicated since the word 'big' is relative, the Big Data concept is clearly about three major characteristics (known as three V's) of data namely: velocity, volume and variety (Zikopoulos and Eaton 2011). While volume relates to the size of data, velocity relates to the data generation speed and the need for analysis of such data, and variety has to do with the extent of variability of data (Zikopoulos and Eaton 2011). Veracity and value have also been recently added to make it five V's (Hitzler and Janowicz 2013). Big data mostly has to do with unstructured data (Suthaharan and Shan 2014). Contrary to widespread perception, however, structured data is also suitable for classification as Big Data as long as the dataset exhibit the necessary features (Zikopoulos and Eaton 2011). There are two main aspects of Big Data, both of which are vital to the proposed solution. They are Big Data Engineering and Big Data Analytics (Figure 4.1).

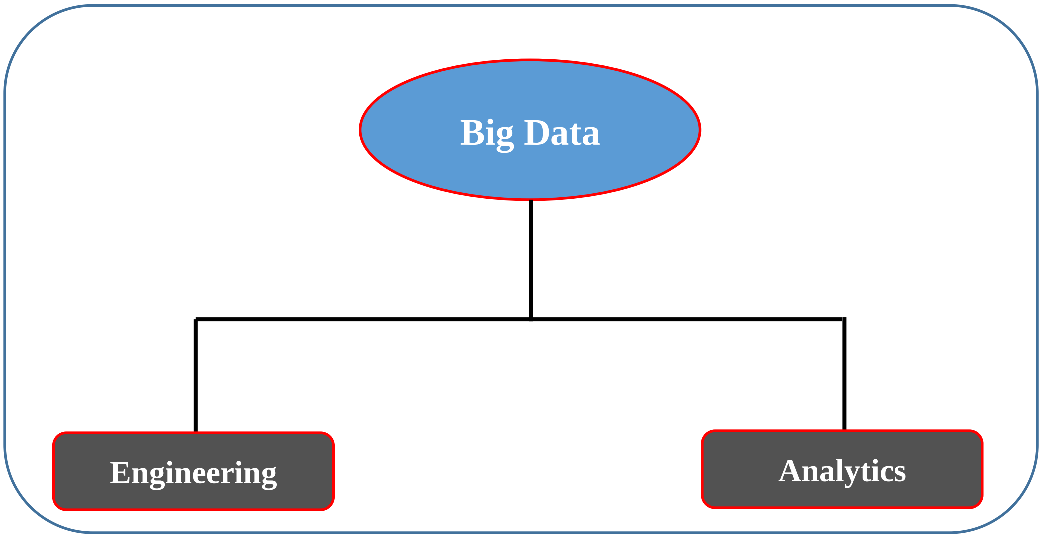


Figure 4.1: Aspects of Big data

4.3 Big data engineering

4.3.1 Hadoop, mapreduce and the problem of classification for prediction

The most common and complete Big Data framework is Apache Hadoop. Apache Hadoop (see Figure 4.2) is a complete open-source Big Data framework for reliable, scalable and distributed computing. It supports processing of huge data distributed across a cluster/assemblage of computers using simple programming model i.e. MapReduce (Hadoop, 2014). Hadoop is fault tolerant and comprises of four modules which are very briefly described after Figure 4.2. The Hadoop Distributed File System (HDFS) and MapReduce are the key components.

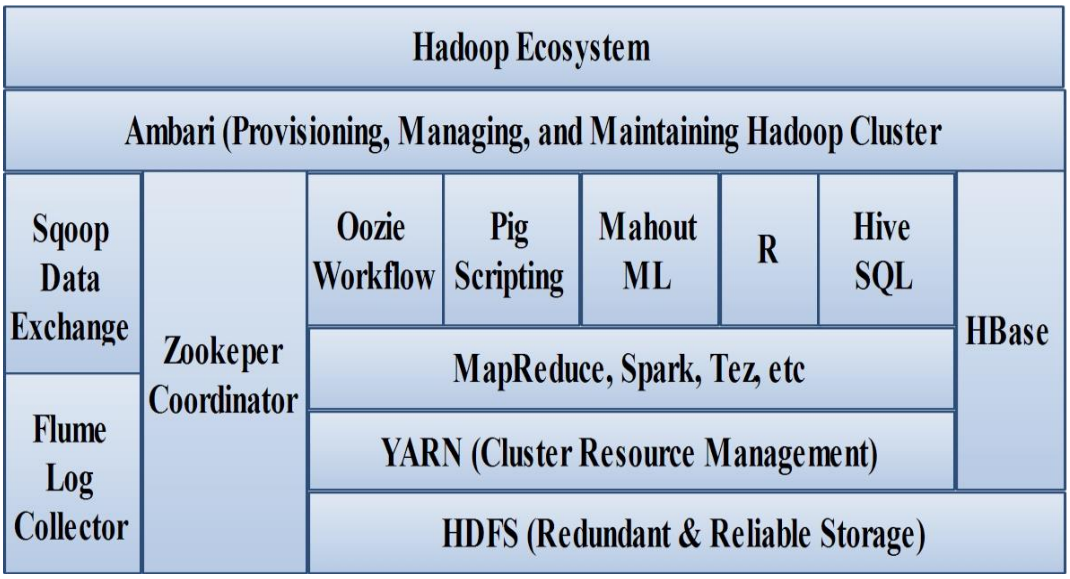

Hadoop Ecosystem
Ambari (Provisioning, Managing, and Maintaining Hadoop Cluster
Sqoop Data Exchange
Flume Log Collector
Zookeper Coordinator
Oozie Workflow
Pig Scripting
Mahout ML
R
Hive SQL
HBase
MapReduce, Spark, Tez, etc
YARN (Cluster Resource Management)
HDFS (Redundant & Reliable Storage)

Figure 4.2: Hadoop Ecosystem

(i) Hadoop common: This encompasses libraries and utilities required by other Hadoop modules

(ii) HDFS: This is a file system that helps to store large data on many machines/nodes and provides direct simultaneous access to such data in a way that is more efficient than reading (the data) from just one machine/node (see Lee *et al.* 2014 for more).

(iii) Hadoop Yarn: This is a framework designed to manage data distribution to nodes during computations (Hadoop, 2014)

(iv) Hadoop MapReduce: This is the computational engine of the Apache Hadoop framework. It is the piece that actually does the action.

> "*MapReduce is a programming model and an associated implementation for processing and generating large data sets. Users specify a map function that processes a key/value pair to generate a set of intermediate key/value pairs, and a reduce function that merges all intermediate values associated with the same intermediate key*" (Dean and Ghemawat 2008: p.107, see this citation for more).

A diagrammatic representation of the MapReduce function is presented in Figure 4.3.

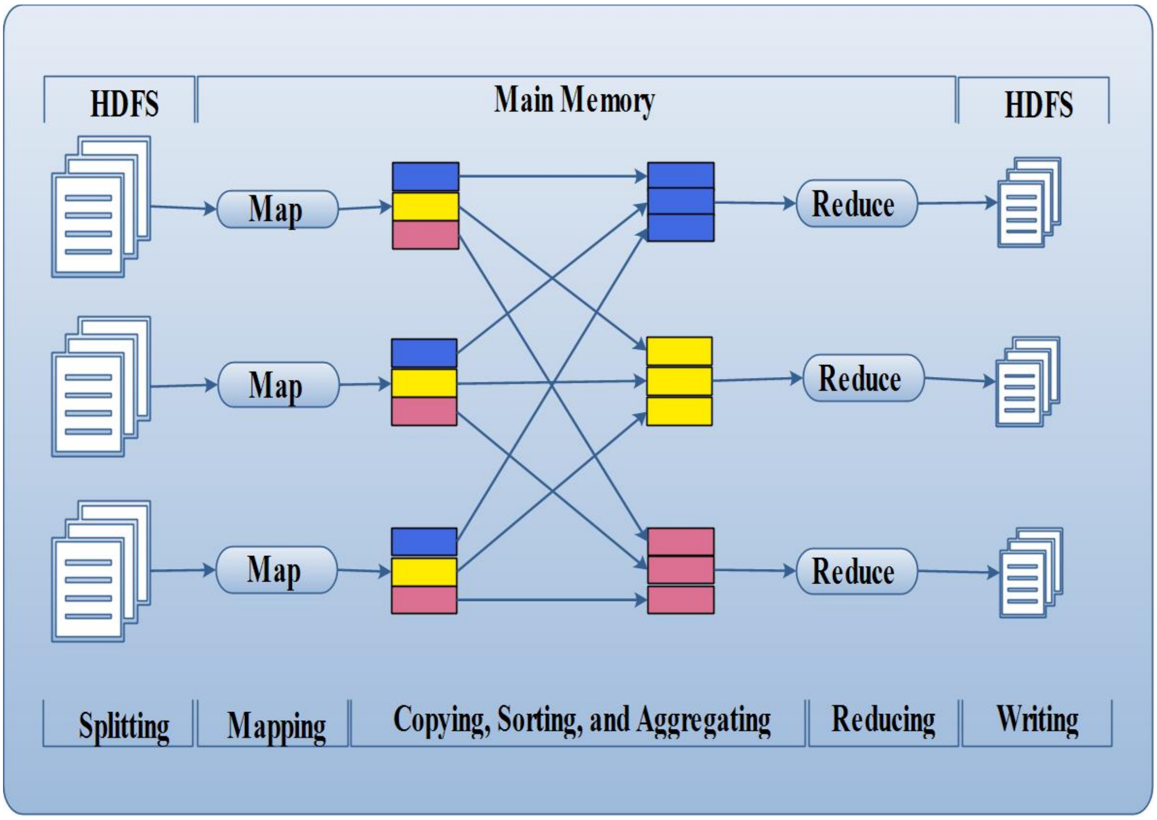


Figure 4.3: How the MapReduce function works

The classification problem of Hadoop: It is well known that artificial intelligence (AI) tools/algorithms formulate classification problems as optimisation problems and tend to perform a large number of iterations which eventually converge on the best solution (Wei and Lin 2010). Unfortunately, MapReduce is very inefficient for iterative data processing since it is only designed to analyse large data, from especially multiple nodes, in one cycle and not repeatedly (Park, 2013) thereby making it unsuitable for the classification analysis required to develop CF-IPM. This has led to the development of some Big Data initiatives that are optimised for iterations (see subsection 4.3.2).

4.3.2 Initiatives for construction firms insolvency prediction on big data platform

None of the tools used for building insolvency prediction models (IPMs) for construction firms has the capability to carry out a robust analysis on any huge data that might require more than a single machine's memory for analysis (Madden, 2012) as required in Big Data Analytics. Using AI tools to analyse Big Data directly is thus virtually impossible (Fan and Bifet 2013). None of the tools can also smoothly operate directly on Hadoop using MapReduce, the most common Big Data platform and the most common execution model respectively. Even when some AI tools operate partially successfully on Hadoop using MapReduce, execution of the much needed iterative computations required for classification when developing CF-IPMs are not achieved (Wei and Lin 2010). This has led to the development of some Big Data initiatives specifically built to support iterative computations which can be used to build CF-IPMs

Presently, there exist some Big Data initiatives developed to optimise the iteration computation aspect of Big Data analysis required for CF-IPM classification of construction firms, most of which are MapReduce-based except one. All these initiatives are open source and can be freely accessed and implemented. The MapReduce-based initiatives include Microsoft's Project Daytona, University of Washington's HaLoop and Indiana University's Twister (Madden, 2012) while the 'Spark' (a relatively new model) based initiative is Apache Mahout which was in itself previously MapReduce-based until April 2014. Another open source initiative currently under development is the Massive Online Analysis (Bifet *et al.*, 2010). Most of these initiatives are discussed in the following sub-subsections. Note that work is ongoing to expand the number of tools supported on all initiatives by their respective developers (Mahout 2015; Ekanayake *et al.* 2010; Barga *et al.* 2012; among others).

4.3.2.1 MapReduce Based Initiatives

MapReduce is the most popular Big Data programming model that currently exist and many of the initiatives are based on it. Most of these initiatives are either in a MapReduce Modified or runtime form.

Haloop*:* This refers to a reworked version of the original Hadoop MapReduce framework to support iteration and some other task/analysis types (Bu *et al.*, 2010). It supports some AI tools (Agneeswaran, 2014) and is okay to develop IPMs for construction firms. Haloop can however only work on clusters and not on a single machine i.e. single node (Bu *et al.*, 2010).

Twister*:* This is a not-heavy MapReduce runtime that enhances the performance of the MapReduce function in a task like iteration (Ekanayake *et al.*, 2010); it is similar to Haloop. The enhancement is mainly in the form of efficiency thereby making MapReduce perform iterations faster. This makes it favourable for developing IPMs for construction firms. Twister works on any shared file system as well as cloud (Ekanayake *et al.* 2010; Zhanquan and Fox 2012) and supports some AI tools.

Microsoft's Project Daytona*:* Daytona is a cloud service iterative MapReduce runtime (Lei *et al.*, 2012)). It is developed to specifically work with Microsoft Azure (previously known as Windows Azure). Microsoft Azure is an open source scalable hybrid-ready cloud computing platform and infrastructure which allows building, deploying and management of applications. Unlike traditional MapReduce runtimes, Daytona maximises the storage services given by Microsoft Azure's cloud platform and infrastructure by using it as the data source as well as data destination (Barga, Ekanayake and Lu, 2012). Daytona supports a wide range of AI tools. The focus on iteration makes Daytona appropriate for developing CF-IPMs. It does not need a distributed file system for implementation

4.3.2.2 Spark Based Initiative

Spark represents an effective alternative to MapReduce. It presents an abstraction known as Resilient Distributed Datasets (RDDs) which efficiently supports artificial intelligence tools executing iterative tools (Zaharia *et al.*, 2010). Unlike traditional MapReduce where the data are often read from and written to distributed file systems, DSS nullifies the need to keep re-uploading data for each iteration thereby greatly increasing performance/efficiency (Zaharia *et al.,* 2010). It allows data to be uploaded onto clusters where they can repeatedly be queried, thereby making the iterative process a lot faster. This is very important for CF-IPMs since the classification process goes through a potentially long iterative process to come up with the optimal equation for the correct/best classification. It is claimed that RDDs enable Spark to outrun MapReduce up to 100 times in multi-pass analytics. The only Spark based initiative is Apache Mahout

Apache Mahout: An open source scalable AI library in Big Data ecosystem. It allows some AI tools/algorithms (e.g. K-means, logistic regression, Naïve Bayes Family, Multilayer Perceptron, among others) to perform classification, clustering and collaborative filtering analysis on large volume data on its scalable distributed file systems or on a single machine (Mahout, 2015). Its codebase, which was previously implemented on Hadoop using MapReduce paradigm, has now been moved to support more execution efficient and richer programming model systems; mainly Apache Spark and H2O (Harris D., 2014). The old Mahout only scaled linearly and supported only AI tools that performed classification by linear analysis e.g. linear SVM (Ericson and Pallickara 2013); it was built on Hadoop and ran only on HDFS.

4.3.3 Initiative selection framework and apache spark as the right initiative for the research

Choosing the right option for developing a Big Data CF-IPM largely depends on the knowledge of the developer and where the construction firms' data are located among other things. Except that it is not MapReduce-based, Apache Mahout with the Apache Spark model is currently the most flexible initiative as it works with almost any database and any distributed file system. The current Apache Mahout is a lot more flexible. It supports over 20 AI tools, whether linear or non-linear, thereby easing the limitation of choice of AI tool for performing analysis on Big Data; it is also fault tolerant, making mistakes identification quite easy during model building (Ericson and Pallickara 2013); It can operate on single or cluster machines inclusive of using cloud as nodes. *Apache Spark is thus the initiative that will be used build the CF-IPM in the proposed solution*.

Other initiatives have certain restrictions that have to be followed for them to be implementable. For example, Microsoft Daytona is only implementable on Microsoft Azure cloud. Table 4.1 provides the features/requirements of each initiative while Figure 4.4 is a framework guide to the selection of initiatives for other users.

Table 4.1: Features of Big data initiatives capable of building CF-IPMs

Big data analytics initiative	**Type/ processing systems**	**Implementation**	**Distributed file system**	**Single or cluster/cloud**	**Support AI tools**	**Fault tolerance (FT)**
Old Apache Mahout	MapReduce	Hadoop platform	HDFS	Both	Yes[+]	Yes
Daytona	MapReduce runtime	Microsoft Azure	Not required[¬]	Cloud-based only	Yes	Yes
Twister	MapReduce runtime	Twister platform	Twister tool	Both	Yes	No[#]
Haloop	Modified Hadoop	Hadoop Platform	HDFS	Cluster/Cloud only	Yes	Yes
Apache Spark	Spark	Any	Any	Both	Yes	Yes

[+] *For linear analysis only*

[¬] *Distributed file system is automatically provided by Microsoft Azure*

[#] *No for iterative computations which are required in classification analysis for IPMs, but is under development (http://www.iterativemapreduce.org/)*

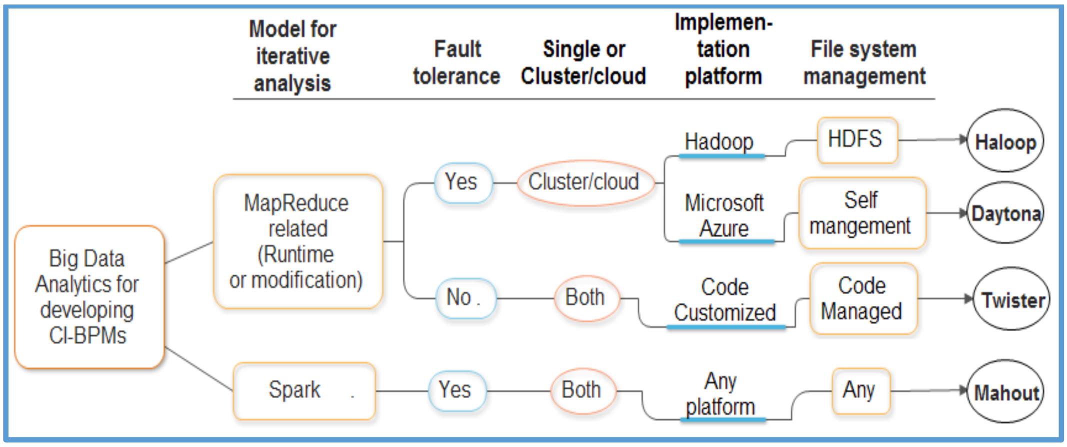


Figure 4.4: A framework for selecting the most suitable Big Data initiative based on developer's skills

Self-management: Platform manages file system

Code managed: Requires an input code on nodes to manage file system

Any (file management system): Supports any file management system

Following the insights offered into the potentials and challenges of CF-IPM development and Big Data Analytics, a framework architecture on how to use each of the Big Data initiatives to develop CF-IPM is given in Figure 4.5. The framework starts with the collection of vast amount of construction firms' data from various financial data sources into a set of computers (commodity servers) before the data is converted to the Key-Value Pair structure compatible with Big Data Analytics platforms/models. Depending on the Big Data initiative adopted, the relevant initiative platform is installed. The corresponding distributed file system is then installed/applied to the data in each of the computers in the commodity server. For example, this can be the installation/application of HDFS for Haloop; code implemented for Twister, or the data simply moved to Microsoft Azure cloud for Microsoft Daytona. With these steps, the Big Data initiative can be executed to carry out the iterative classification analysis required for developing the construction firms Big Data IPM.

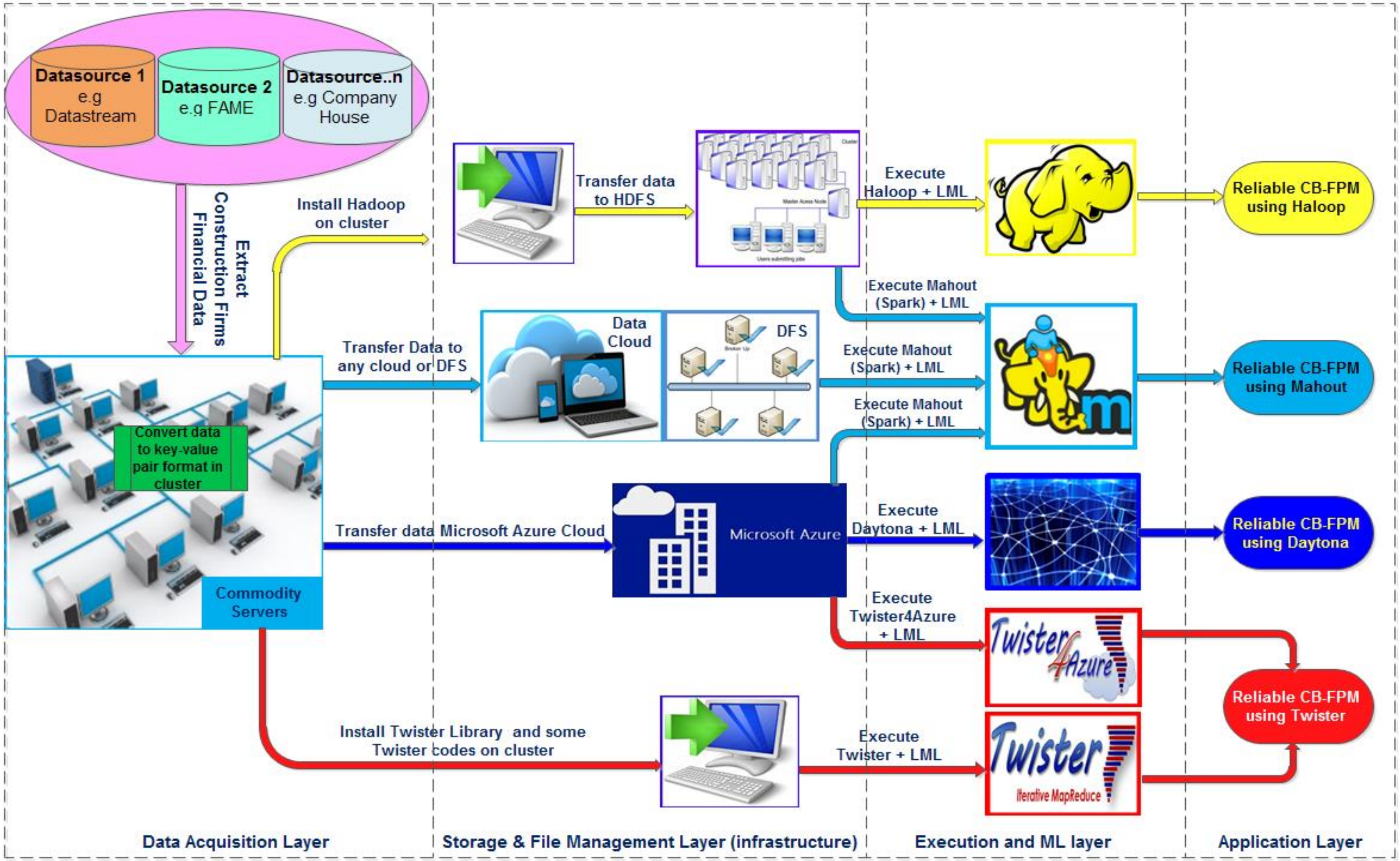


Figure 4.5: Framework architecture for developing highly reliable construction firms IPM or CF-IPM using Big Data

Note: DFS = Distributed File system

4.4 Big data analytics

With robust analytical and data mining capabilities, Big Data conducts advanced analytics such as predictive analytics, inferential analytics, prescriptive analytics and descriptive Analytics (Ohlhorst, 2013; Talia, 2013), among others. The type of analytics used in the proposed solution is predictive analytics hence will be described first ahead of other types. Table 4.2 presents some common iterative classifications tools and the Big Data Analytics initiatives that can implement them according to literature.

Table 4.2: Some artificial intelligence classification tools supported by Big Data initiatives

Tool	Supporting Big Data Initiatives
Logistic regression or Logit Analysis	Mahout (Mahout, 2015); Daytona (Barga, Ekanayake and Lu, 2012)
Multi-discriminant analysis	Mahout (Mahout, 2015); Daytona (Barga, Ekanayake and Lu, 2012)
Random Forest	Mahout (Mahout, 2015); Daytona (Barga, Ekanayake and Lu, 2012)
Support vector machines	Mahout (Mahout, 2015); Haloop (Bu *et al.*, 2010); Twister (Zhanquan and Fox 2012)
Naïve Bayes or Bayesian Classifier	Mahout; Daytona (Barga, Ekanayake and Lu, 2012)
Artificial neural networks	Haloop (Bu *et al.*, 2010); Daytona (Barga, Ekanayake and Lu, 2012); Mahout (Mahout, 2015); Twister (Gu, Shen and Huang, 2013)

Predictive analytics: Prescriptive analytics is concerned with the prediction of future probabilities, trends and patterns within a dataset based on past happenings (Sagiroglu and Sinanc 2013). This is the aim of this study as given in section 1.7. Prescriptive analytics answers the question: what will happen?

Inferential analytics: Inferential analytics sometimes taken as a subordinate of predictive analytics. It focuses on the interactions of explanatory variables with the target variable in the dataset (LaValle, Lesser and Shockley, 2011). It is mainly used to check the independent variables that have the most impact on the target, and the type of relationship that exist them. This form of analytics is partly used in the proposed solution; it is used for the variable selection process

Descriptive analytics: Descriptive analytics is used to examine what particular event happened, usually using real-time data (Xindong Wu *et al.*, 2014). It is used to answer the question: what happened?

Diagnostic Analysis: Diagnostic analysis, as implied by its name, is used to diagnose an event by checking what led to the event. It is used to answer the question: why did it happen?

Prescriptive analytics: Prescriptive analytics involves using optimisation and simulation algorithms to propose possible outcomes and answers (Boyd and Crawford 2012) to problems. It is used to answer the question: what is the best course of action?

4.5 Justification of use of big data analytics for the construction firms insolvency prediction data to be used

A dataset can be taken to be Big Data when its velocity, volume variety and (or) veracity become so much that current technological tools make a harsh of storing or processing it (Pflugfelder and Helmut 2013; Suthaharan and Shan 2014). Its size is such that it forces a search for new approaches away from the known and trusted ones. In the past, say around the 80s, it would have been a data size that required 'tape monkeys'; presently, it is a data size that will require clusters of computer and/or cloud running concurrently and in a parallel mode to be analysed (Fan and Bifet 2013). Big Data Analytics can be defined as involving analysis of huge data in order to unmask valuable patterns/information (Suthaharan and Shan 2014).

Although size is a key feature in qualifying data as 'Big Data', the nature of the analysis is as important as much. Jacobs and Adam (2009), in his experiment, showed why a dataset could qualify or not qualify to be classified as Big Data. Jacobs and Adam (2009) created a demographic data (religion, marital status, ethnicity, among others) of the world population in a table of circa ten columns and over 7 billion rows which were contained in a 100-gigabyte hard disk. Simple programs written to return answers to queries like the mean age of the world population ran smoothly on a computer with low-performance CPU, thus not making the data viable to be classified as Big Data. An attempt to simply load the same data, without performing any analysis, on a commonly used enterprise-grade database system (PostgreSQL6) running on a super performance computer (an eight-core Mac Pro workstation equipped with 20 gigabyte RAM and two terabytes of RAID 0 disk) had to be aborted after six hours of unsuccessful upload. A serious analysis of the created data on this database will obviously take days if not weeks or months hence it can be classified as Big Data, in this case, based mainly on analysis.

The above example is what makes the data of hundreds of construction firms in a country over some years qualify as Big Data. A simple input of such data into columns and rows of Microsoft Excel and finding averages might not be considered as ‘Big’ in the present technological world. However, a more complex analysis (like iterative classification analysis which is used for insolvency prediction) of such data using a high computational demand AI tool will be very onerous on any computer. Such analysis hence qualifies the data for Big Data Analytics. The best example of this is a study by Du Jardin (2010) which used 500 firms with well tuned AI tool (artificial neural networks in this case) parameters. The effect of this was a very good model and tedious computational duration of five days with 30 PCs running Windows. With contemporary technology like Big Data Analytics which could do the same computation in seconds, there is no reason why large data and well-tuned AI tools should be avoided any longer in CF-IPM. The proposed solution will thus use a relatively large data (about a thousand construction firms), well-tuned and powerful AI tools, and Big Data Analytics.

4.6 Chapter summary

This chapter introduced Big Data and explained how it will be used to build a CF-IPM in the research. The general definition of Big Data is data with high velocity, volume and variety (as well as veracity and value) which the present technology cannot comfortably analyse. The most common Big Data framework is Apache Hadoop and the most common programming model or executioner is MapReduce.

To build a CF-IPM, AI tools formulate classification problems as optimisation models and tend to perform a large number of iterations. MapReduce is however very inefficient for iterative data processing. However, on the positive side, there are some initiatives that have been developed to solve this problem, as presented in this chapter, including Microsoft's Project Daytona, University of Washington's HaLoop, Indiana University's Twister and Apache Mahout which currently uses University of California’s Spark. Compared to others, Apache Mahout (Spark) is a lot more flexible and supports over 20 AI tools, whether linear or non-linear, and is thus the initiative that will be used build the CF-IPM in the research. Regarding Big Data Analytics, the proposed solution will be using predictive analytics since the future probability of failure of construction firms will be predicted using past (from last year) data of existing and failed construction firms

The suitability of about a thousand construction firms, which is what will be used as data in the proposed solution, for Big Data Analytics was also justified in this chapter. For example, Du Jardin (2010) used 500 firms with a well tuned AI tool, taking a duration of five days with 30 PCs running Windows. Over 500 firms (about a thousand), which is the sample size for the proposed solution, with well-tuned AI tools will take longer and (or) more PCs hence the need to use Big Data Analytics in the proposed solution, which could do this tedious computation in minutes.

A comprehensive review of CF-IPM journal articles with focus on methodical issues that have led to poor performing CF-IPMs were presented in chapter five.

CHAPTER FIVE

5.0 METHODICAL ISSUES IN INSOLVENCY PREDICTION MODELS FOR CONSTRUCTION FIRMS

5.1 Chapter introduction

The methods used to develop a construction firms insolvency prediction models (CF-IPM) can make or break the model and have long been scrutinised by researchers. Many studies have ignorantly used invalid, less rigorous or questionable methods to develop their CF-IPMs simply by copying earlier studies. There are many studies that have in fact neglected the responsibility of justifying their methods. This chapter is a review of the methods involved in developing construction firms insolvency prediction models (CF-IPM). It deals with the types of tools, sample characteristics and model testing features of CF-IPM studies. In each case, the model feature being analysed is briefly explained, the trend in CF-IPM is presented in a table, and (or) charts, and the most effective and (or) poorest ways of implementing the features are discussed along with the CF-IPM trend. After this, a statement on how the best practice of implementing the feature will be adopted in the proposed solution will be made. Thus the final paragraph in each sub-section could as well be the most important (especially for a quick read).

The next section (i.e. section 5.2) uses the ‘systematic review of construction firms insolvency prediction model studies’ presented in section 3.3 of chapter three but excludes the Hall's (1994) study for its unclear methods. However, a new summary of findings table on the methodical features of the studies is presented here. All the work in this chapter is based on this systematic review and a new summary of findings table. Subsection 5.2.1 is a review of the trend on the types of tools that are used in CF-IPM research and what tools have been ignorantly left out. Subsection 5.2.2 is a review of the data characteristics of the data used to build CF-IPMs regarding dispersion and explains what is best to be used. Subsection 5.2.3 is an examination of how CF-IPM studies have validated their models over the years with a view to identifying best practice. Subsection 5.2.4 is a scrutiny of CF-IPM studies on their consideration of error cost, or otherwise, and explains the best practice on this. Section 5.3 is a presentation of the summary of this chapter.

5.2 Systematic review of construction firms insolvency prediction model studies with focus on methods

The method used to carry out the systematic review in this section is the same as that presented in section 3.3. However, Hall's (1994) study was removed from the 31 studies reviewed in this section because it was not clear about its methods. In essence, only 30 of the 31 systematically reviewed studies in section 3.3 are examined here. To assess the methods of the 30 primary studies (i.e. the systematically reviewed studies), the tools used, characteristics of data or sample used, the process of validating the model, and the consideration of error cost of the model in the studies were assessed. The trends were checked against what is more fitting for construction firms and (or) what is more common in the insolvency prediction model (IPM) literature. As required for systematic review, a meta-analysis based the methods used in the studies was done with data synthesised through the use of 'Summary of Findings' tables (Higgins and Green 2008; Smith *et al.* 2011) in Table 5.1.

5.2.1 Tools used (accuracy and transparency)

The tool used to build a construction firms insolvency prediction model (CF-IPM) plays a big role in the performance of such model. The norm is to use a number of tools to build a model and compare so as to select the one that produces the better model. The two main categories of tools are the statistical and artificial intelligence tools, although there is also the uncommon option-based model formula. The most popular statistical tools as noted by Balcaen and Ooghe (2006) in their comprehensive review of insolvency prediction models (IPMs) are multiple discriminant analysis (MDA) and Logistic regression (LR). The most popular artificial intelligence (AI) tools as advocated by (Adnan Aziz and Dar 2006; Ravi Kumar and Ravi 2007) in their comprehensive reviews are artificial neural network (ANN), support vector machines (SVM), genetic algorithm (GA) and decision tree (DT).

Table 5.1: Summary of finding showing data, tools and model testing related features of the primary studies

	Author (Year)	Tools used	Class balancing method	Variable selection method	Authors background	Accuracy	Error type considered	Total sample	Existing firms (%)	Failed firms (%)	Validation	Sample % used for validation	Var. type and no.	Period
1.	Fadel (1977)	LR	LR	SPR	AEF	-	-	102	-	-	N	-	2 FR	68-73
2.	Mason and Harris (1979)	MDA		SP	CE	87	-	40	50	50	11 firms		6 FR	69-77
3.	Kangari and Farid (1992)	LR	-	Lit rev	CE	-	-	126	-	-	-	-	6FR	82-88
4.	Langford *et al.* (1993)	MDA		Lit rev	CE; BD	63.33	N	3	0	100	Y	100	5 FR	90-93
5.	Abidali and Harris (1995)	MDA and A-score	UB	SPDA	CE	70.3	Y	112	19.6	80.4	HOV	72.3	7 FR	78-86
6.	Russell and Zhai (1996)	Discriminant function using SPR	UB	SPR	CE	78.3		143	58.7	41.3	HOV	16	6 Fin vars	75-93
7.	Koksal and Arditi (2004)	FA; LR	UB		CE	-	-	53	79.2	20.8	R^2 value	-	21 non-Fin vars	
8.	Singh and Tiong (2006)	MCDM	-	Lit rev	-		-	-	-	-	5 firms		5 FR	

	Author (Year)	Tools used	Class balancing method	Variable selection method	Authors background	Accuracy	Error type considered	Total sample	Existing firms (%)	Failed firms (%)	Validation	Sample % used for validation	Var. type and no.	Period
9.	Chen (2009)	ordinary least-squares		SPR	BM	86.13	-	42			HOV	30		97-06
10.	Huang (2009)	Structural model and LR Univariate	UB	LR	CE	88.5	ROC	40	75	25	-	-	4FRand 3MVR	99-06
11.	Sueyoshi and Goto (2009)	DEA–DA; PA; LR	DEA feature	Lit rev	IM	93.9	N	215	90.7	9.3			5FR	98-05
12.	Stroe and Bărbuță-Mișu (2010)	MDA	-	-	BM	77.8	-	11	-	-	10 firms	-	5 FR	01-06
13.	De Andrés *et al.* (2011)	SOM and MARS hybrid.		SPDA	AEF and ME	88.72	Y	63107	99.6	0.4	CV	20	5FR	07-08
14.	Ng *et al.* (2011)	MDA	UB	SP	CE	96.9	N	35	88.5	11.5	N	-	7 FR	
15.	Tserng *et al.* (2011)	BSM; CB; BS	USP	OMV	CE and AEF	90%	ROC	87	66.7	33.3	Y	-	SM	70-06

	Author (Year)	**Tools used**	**Class balancing method**	**Variable selection method**	**Authors background**	**Accuracy**	**Error type considered**	**Total sample**	**Existing firms (%)**	**Failed firms (%)**	**Validation**	**Sample % used for validation**	**Var. type and no.**	**Period**
16.	Tserng *et al.* (2012)	Barrier option model; MDA	UB	OMV	CE and AEF	84.5	ROC	121	76	24	CV	-	6OMVand 4FR	70-06
17.	Tserng, Lin, *et al.* (2011)	ESVM; LR	OSP	SPDA	CE	80.31	ROC	168	69.6	30.4			7 FR	
18.	Chen (2012)	Hybrid SFNN	UB	All in FS	CE	85.1	-	42	35	65	CV	10	25 FR	98-08
19.	Sánchez-Lasheras *et al.* (2012)	SOM and MARS hybrid.	UB	SPDA	CE, ME and AEF	89.58	Y	63107	99.6	0.4	CV	20	5 FR	
20.	Tsai *et al.* (2012)	BSM and LR hybrid	UB	SPR	CE and AEF	87.32		121	76	24	CV		4FR and 1OMV	70-06
21.	Horta and Camanho (2013)	SVM	USP and OSP	Lit rev	Engineering	97.6	ROC	10559	85	15	HOV	20	6FR and 3Strat. var	08-10
22.	Makeeva and Neretina (2013a)	MDA; LR; PA	BA	FAand SPR	AEF	86.44	Y	120	50	50	-	-	6FR	02-10

	Author (Year)	**Tools used**	**Class balancing method**	**Variable selection method**	**Authors background**	**Accuracy**	**Error type considered**	**Total sample**	**Existing firms (%)**	**Failed firms (%)**	**Validation**	**Sample % used for validation**	**Var. type and no.**	**Period**
23.	Makeeva and Neretina (2013b)	MDA; LR; PA	USP	FAand SPR	AEF	86.44	Y	120	50	50	-	-	22 FR	02-10
24.	Sun *et al.* (2013)	ANN-AB hybrid; ANN bagging hybrid; ANN	USP	t-test, CA and SPDA	AEF	93.07	6.93	85	61.2	38.8	CV	33.3	9 FR	01-10
25.	Cheng *et al.* (2014)	LS-SVM and DE hybrid; SVM; ANN	OSP by SMOTE	SPDA	CE	92.13	ROC	76	82.9	17.1	CV	20	7 FR	70-11
26.	Heo and Yang (2014)	AB; DT; ANN; SVM; MDA	USP	Lit rev	CoE	78.5	Y	2762	50	50	HOV	20	5FR	08-12
27.	Muscettola (2014)	LR	UB	SPR	AEF	80.94	Y	1338	87.2	12.8	HOV	-	9FR	07-11
28.	Tserng *et al.* (2014)	Univariate and LR multivariate	UB	LR	CE	79.18	ROC	87	66.7	33.3	CV	-	4FR and 1MVR	70-06
29.	Cheng and Hoang (2015)	KNNand FFA hybrid, SVM; MDA; LR	OSP	Lit rev	CE	96.0	ROC	76	82.9	17.1	CV	20	20 FR	70–11

	Author (Year)	Tools used	Class balancing method	Variable selection method	Authors background	Accuracy	Error type considered	Total sample	Existing firms (%)	Failed firms (%)	Validation	Sample % used for validation	Var. type and no.	Period
30.	Tserng *et al.* (2015)	GST	-	GST	CE	84.8	-	92	73.9	26.1	-	-	8 FR	72-08

-: Not stated or not clear or not done or not applicable AB: AdaBoost DT: Decision tree AEF: Accounting and/or Finance/ and/or Economics ANN: Artificial neural network BD: Building Department BSM: Black-Scholes-Merton

BA: balanced or equally dispersed data BM: Business Management or Business Administration BS: Bharath and Shumway naïve model CA: Correlation analysis CB: Crosbie and Bohn model CE: Civil Engineering or Construction Engineering CoE: Computer engineering CV: Cross-validation DEA–DA: Data Envelopment Analysis–Discriminant Analysis ESSVM: enforced support vector machine FA: Factor analysis FFA: firefly algorithm FR: financial ratio FS: financial statements GST: Grey system theory HOV: Holdout validation IM: Information ManagementKNN: K-nearest neighbour Lit rev: Literature review LR: Logistic regression MARS: Multivariate Adaptive Regression Splines.

MCDM: Multiple-criteria decision-making MDA: Multiple discriminant analysis ME: Manufacturing engineering Nikkei Needs Corporate Financial Database MVR: market value ratio N: No OMV: Option model variables

OSP: Oversampling Probit analysis: PA ROC: Error cost considered using Receiver Operating Characteristic Curve

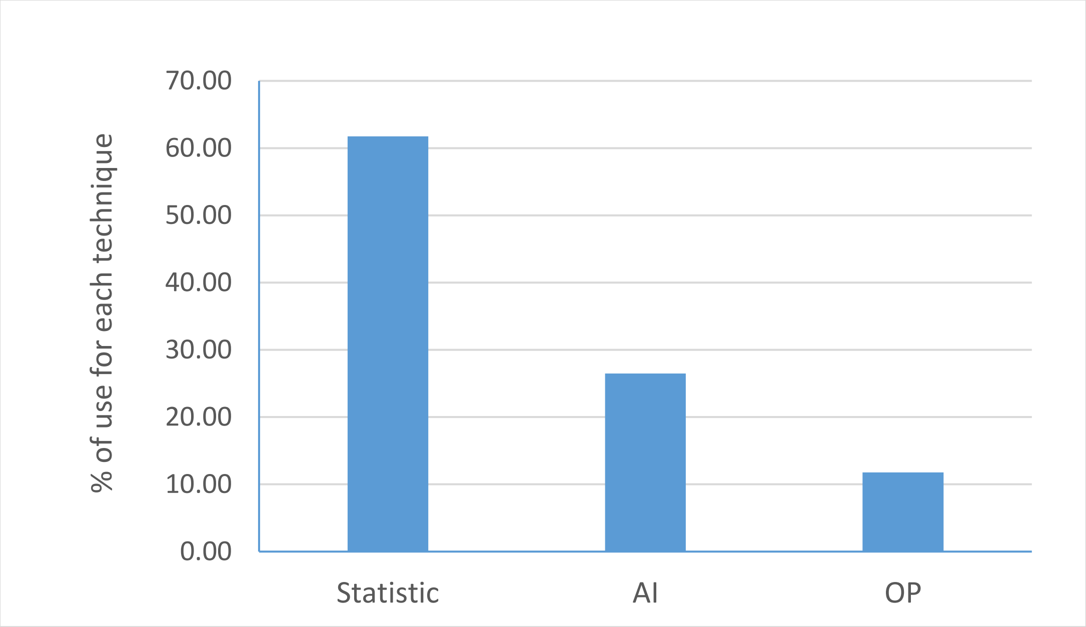
SFNN: self-organizing feature map optimisation, fuzzy, and hyper-rectangular composite Neural Networks SM: Stock market variables

SMOTE: Synthetic Minority Over-sampling Technique SOM: Self-Organizing Maps Neural Networks SP: Stepwise SPDA: discriminant analysis

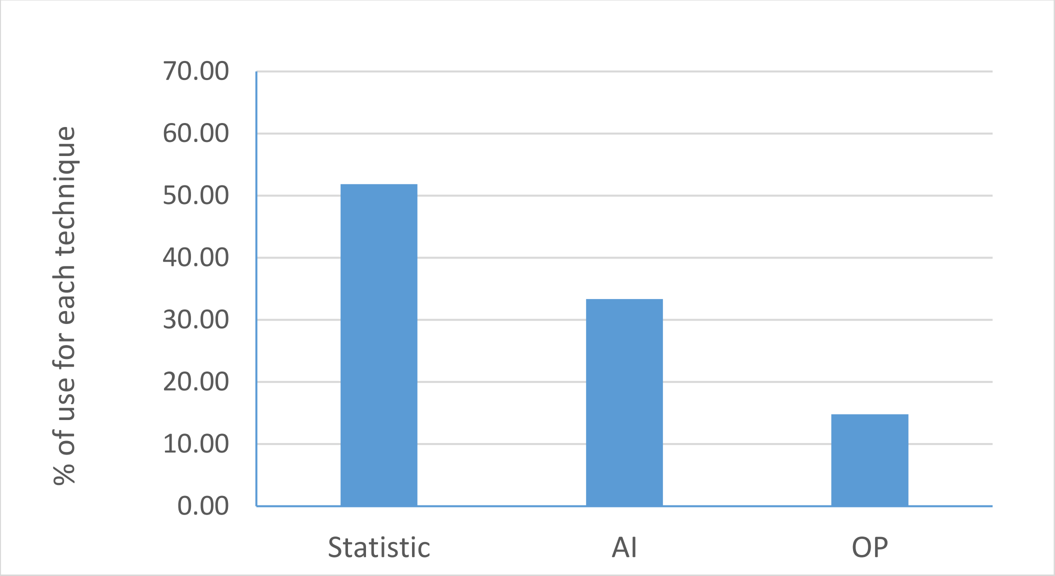
SPR: Stepwise regression Strat. Var: Strategic Variables SVM: Support vector machine UB: unbalanced data USP: Undersampling Y: Yes

Figure 5.1a, b and c contain statistical charts of used tools. The figures consider each tool category used as one study hence if a study used both statistical and AI tools, it is considered it as two studies. From Table 5.1 and Figure 5.1a, it is clear that CF-IPM studies have used more statistical tools from inception till date. This is not surprising as very early works simply followed in the footsteps of Altman (1968) and Ohlson (1980) who pioneered MDA and LR respectively in the IPM research area. Figure 5.1b is a chart of relatively recent works from 2006, when Balcaen and Ooghe (2006) clearly identified the rising popularity of AI tools in the IPM world, until present. The figure portrays that the CF-IPM studies have not adequately adopted the use of AI tools despite the many disadvantages of statistical tools.

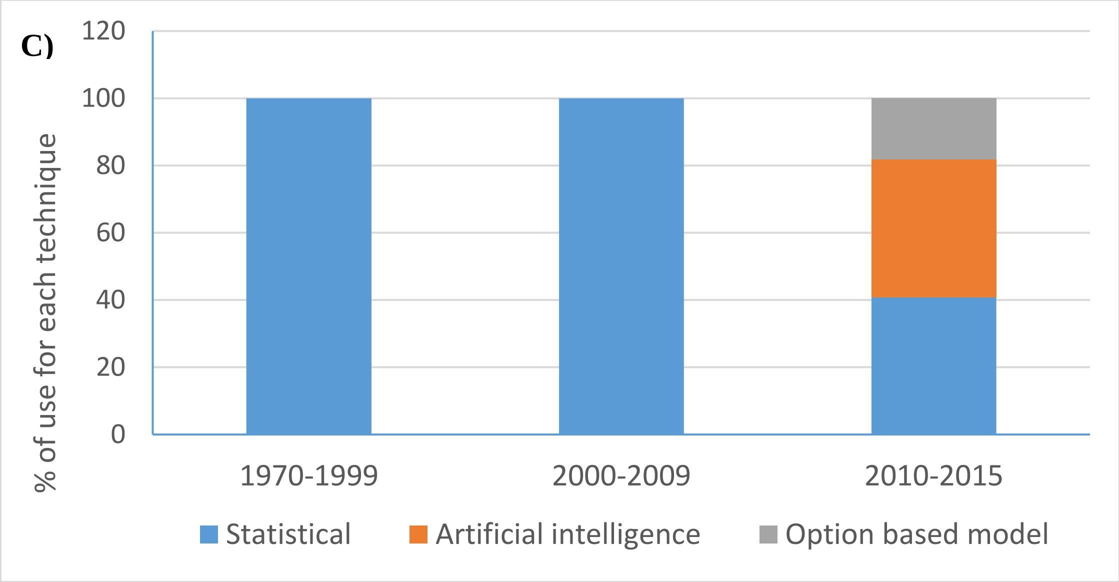
Considering that the use of AI tools in corporate insolvency prediction started as far back as the early 80s (Tserng, Lin *et al.* 2011), Figure 5.1c shows that CF-IPM studies were clearly too slow to take up AI tools. In fact, CF-IPM studies only started using AI tools and option models in or after 2010. However, AI tools increased frequency of use in more recent times (i.e. from 2010 to 2015) is encouraging but should improve because of the many disadvantages of statistical tools and advantages of AI tools, though some studies only employ statistical tools for comparison purpose (e.g. Heo and Yang 2014; Cheng and Hoang 2015).



a) Frequency of tools used in CF-IPM since inception



b) Frequency of tools used in CF-IPM since 2006



c) Frequency of tools used in in landmark periods

Figure 5.1: The frequency of use of tools in the CF-IPM research area

AI: artificial intelligence OP: option-based model

The many disadvantages of statistical tools, which have been identified in many studies over the years, are normally in form of restrictive assumptions that data need to satisfy for statistical tools to perform optimally (Altman 1993; Balcaen and Ooghe 2006; Chen 2009; du Jardin and Séverin 2011; Joy *et al.* 1975; among others). Some of these assumptions include: (i) that independent variables must have multivariate normality, (ii) that each group data (i.e. failed and existing firms' data) must have equal variance-covariance, (iii) that groups must be discrete and non-overlapping, among others (Balcaen and Ooghe 2006;

Sueyoshi and Goto 2009; Ng *et al.* 2011; Tserng *et al.* 2015). All these restrictive assumptions can barely be satisfied together by one data set hence are violated in part or in totality in all cases i.e. in many studies (Richardson and Davidson 1984; Zavgren 1985; Chung *et al.* 2008). Nonetheless, LR is deemed relatively less demanding compared to MDA (Altman 1993; Balcaen and Ooghe 2006; Jackson and Wood 2013) hence the high use of LR compared to MDA and probit analysis (PA) in CF-IPM studies (Figure 5.2) is commendable.

It is accepted and well proven in many IPM studies that AI tools produce better IPMs (Chen 2012; Divsalar *et al.* 2012; Heo and Yang 2014; Tserng, Lin *et al.* 2011; Yoon and Kwon 2010; Zhou, Lai, and Yen 2014; among many others). A major reason for the slow take-up of AI tools in CF-IPM studies could be as a result of the reluctance of construction academics, who are responsible for two third of the primary studies (see Figure 5.3), to learn how to use them as they usually require some level of computing skills. An increase in the use of AI tools will ensure better models are built in CF-IPM studies.

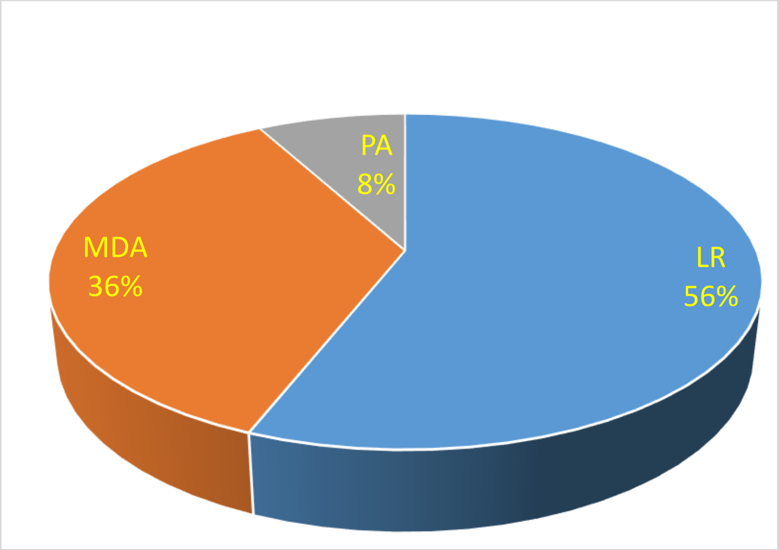


Figure 5.2: Frequency of use of statistical tools in the CI IPM study area

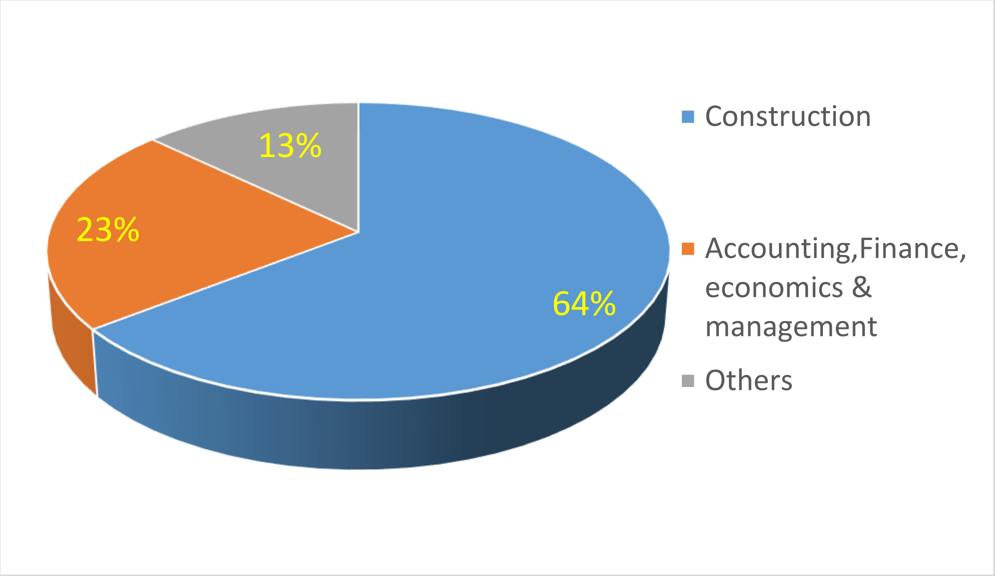


Figure 5.3: Proportion of CF-IPM studies by authors' background

Overall, the more contemporary issue with the CF-IPM studies is the limitation of use of AI tools to mainly ANN, SVM and DT, with the exemption of few studies like Sun *et al.* (2013) that used Bagging and Adaboost, Heo and Yang (2014) that used Adaboost and Cheng and Hoang (2015) that used KNN. There needs for increased use of AI tools like Adaboost, bagging and KNN to allow better comparison selection of the best performing models based on tools used. There is also need to adopt and compare (models created by) other high performing AI tools like random forest, Adabag Boosting, Extremely Randomized Trees, Naive Bayes, Clustered Support Vector Machines, among others. This will ensure the very best model is selected from the multitude produced from these high performing tools. This is hence the method that will be adopted in this work.

5.2.2 Data or sample characteristics

Data or sample characteristics are very important to the performance of CF-IPMs or even any IPM at all. Data dispersion, defined as the ratio of failing to non-failing (or existing) construction firms or vice versa in a sample data, plays a significant role in building a CF-IPM. Data with equal or near equal dispersion between failing and non-failing construction firms is the very perfect type of data for the optimal performance of any tool. However, the relatively high number of existing construction firms compared to failed firms means that data available to build CF-IPMs are normally highly skewed, a situation which drastically reduces the predictive performance of virtually all tools, especially the statistical ones (Boritz *et al.* 1995; Balcaen and Ooghe 2006). According to du Jardin (2015), highly skewed data normally means that "data that characterised failed firms would be hidden by those that represent non-failed firms, and therefore would become rather useless" (p.291) hence it is best to have equal dispersion (Jo, Han and Lee, 1997). This problem has long been identified and some techniques have been proposed as a solution:

- ❖ Tool's balancing feature: the process whereby the tool employed to develop the model has a special feature that is used to balance/equalise the data dispersion.
- ❖ Over (under) sampling: the process whereby the smaller (larger) group is increased (decreased) until the number of construction firms in it equal that of the larger (smaller) group. The increment in oversampling is usually done by using average values of variables of the firms in the smaller group to form data of new fictional firms for the group until it has equal (or almost equal) number of firms with the larger group. The decrement in undersampling is done by matching firms with similar properties (size, turnover, among others) from the larger group to those of the smaller group until all the firms in the smaller group have a pair in the larger group, then the excess in the larger group is discarded.

Of the 30 primary studies reviewed, only 25 studies clearly presented the level of dispersion of data, showing that not all studies recognise the importance of the data dispersion characteristic. The unrecognised studies include some that are published as recent as in the 2000s (see Table 5.1). Of the 25 that clearly presented data details, less than a fifth (or four studies) used equal data dispersion, about a quarter used some form of data balancing while more than half used unequal data dispersion (Figure 5.4). This problem does not appear to be time-related as it has been highlighted since the pioneering days of IPM studies (Altman, 1968; Ohlson, 1980; Boritz, Kennedy and Albuquerque, 1995). More so, only two of the primary studies reviewed were published before 1990 and a total of six before 2004 (see Table 5.1) making 24 primary studies, or 80% of the reviewed studies, relatively recent.

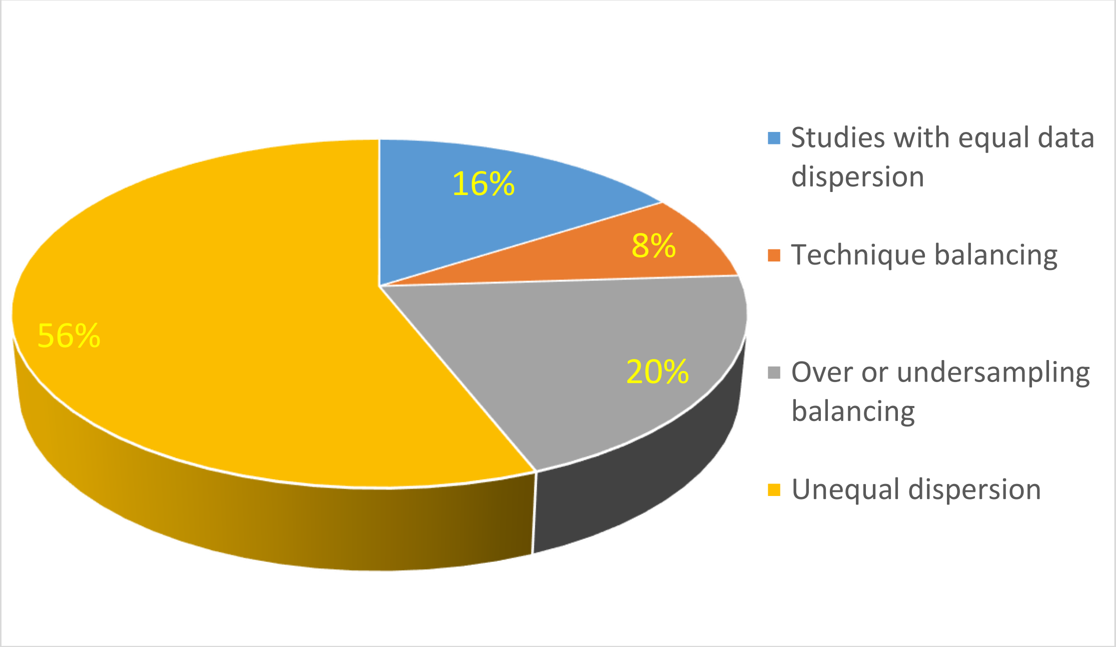


Figure 5.4: Percentage of studies with equal, modified equal and unequal data dispersions

Critics of undersampling are sceptical that discarded data might be those that are crucial to the learning/development process of any tool (Cheng and Hoang 2015). Critics of data balancing, in general, have also argued that it leads to sampling bias and thus the entire population should be used (Agarwal and Taffler 2008; Tserng *et al.* 2012). However, it is well established that using skewed data results in the model being more accurate for predicting the larger group (Boritz *et al.* 1995; Sueyoshi and Goto 2009; Ng *et al.* 2011; du Jardin 2015). This means the model will be more likely to predict an insolvent firm as being solvent than vice versa incorrectly; this is the costlier of the two IPM error types (**see section 5.3.4)** and needs to be well avoided hence equal data dispersion is more appropriate for developing CF-IPMs. The proposed solution will hence use data with equal or almost equal dispersion, employing the undersampling technique with matched samples. With more than half CF-IPM studies using unequal data dispersion, many CF-IPMs must have been suboptimal.

5.2.3 Model validation

In the early days of IPM research (Altman, 1968; Ohlson, 1980), it was common to test a developed model with the data used to build the model. Such tests yielded very accurate results. However, with the models not performing as well in practice, further research by

Joy *et al.* (1975) quickly revealed that IPM developers confused ex-post classification results with ex-ante predictive abilities. Some studies (Joy *et al.* 1975; Taffler 1983; among others) hence rightly recommended that a built model should be tested on separate data apart from that which was used to build it if there is to be any confidence in the model. This practice has now almost become a norm in the IPM research world with the separate data usually referred to as test or validation data.

Normally the data is pre-divided, usually in a ratio of between 80-20 or 70-30, the bigger portion used for training the model and the smaller used for validation. The case where the validation (or test) samples are removed in batches such that the entire sample, at different times, form part of the training or validation sample is known as cross-validation.

The research reveals the relatively poor trend of CF-IPM studies where less than two third of them validated their model (Figure 5.5). Although Ng *et al.* (2011), authors of one of the primary studies, claimed to have validated their model, the research disagrees with that claim because the validation was done using earlier years' data of the sample used to build the CF-IPM. Unfortunately, the immediate earlier year's data of a model building sample is not an acceptable replacement for data of firms that were not used in the model building process. No wonder the model misclassified only one firm out of the 32 firms selected from the model building sample to validate the firms. This single misclassification was even put down to unequal data dispersion by Ng *et al.* (2011) as they unjustifiably tried to explain the perfection of the model.

Of the 11 studies that did not test/validate their models, two did not report a clear accuracy result while eight of the remaining nine reported accuracy values of over 80% (Figure 5.6), depicting highly accurate models. Such accuracy values are clearly unsatisfactory and unacceptable at the very least. In fact, with AI tools like ANN and SVM, it is possible to build a model with a 100% prediction accuracy when tested on training (i.e. model building) data. However, such models are not usually very good on test/validation data and are normally condemned for what is known as 'overfitting' to the training data; this makes such models rather poor (Ahn, Cho, and Kim 2000; Ravi Kumar and Ravi 2007; Tseng and Hu 2010; among others).

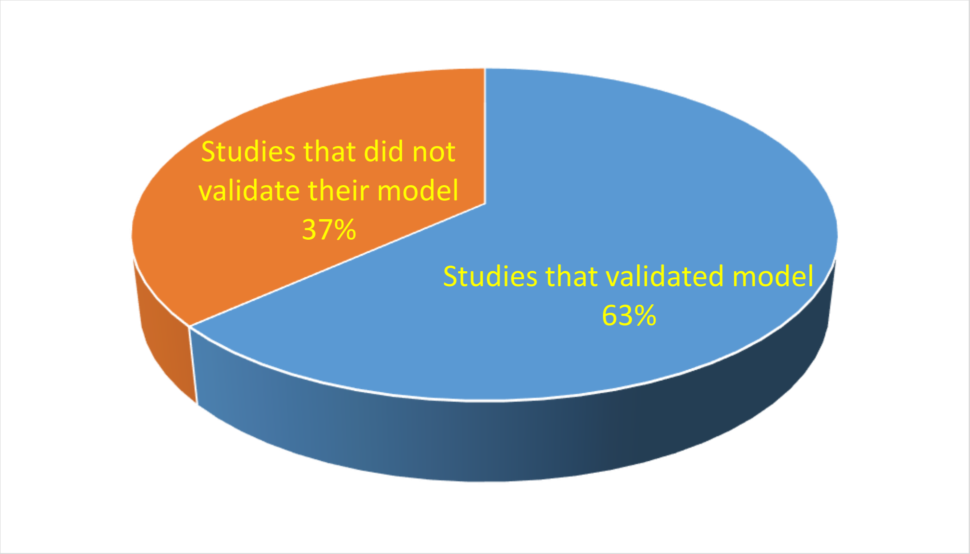


Figure 5.5: Percentage of studies that validated or did not validate their model

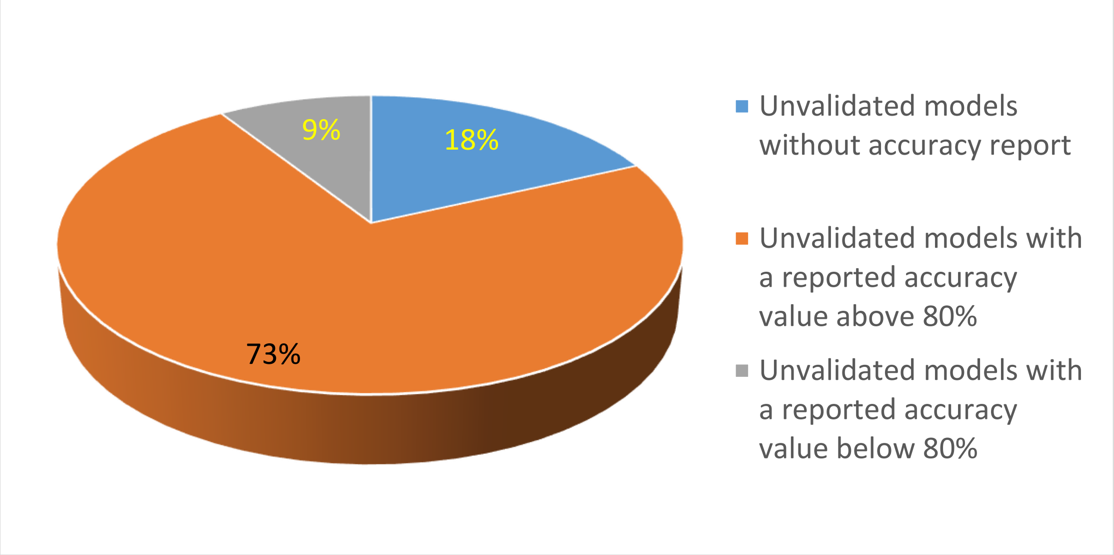


Figure 5.6: Accuracy value range of studies with validated and un-validate their models

In essence, the true accuracy or general performance of a model can only be assessed using a separate data from the one the model was trained or built with. This is because the actual users, which are construction firms in this case, will be using the data of their firms which does not constitute part of the model building data; and will be expecting to get a reliable result about the status (failing or healthy) of their construction firms. The approach in the proposed solution is thus to set aside a minimum of 20% of the data for testing any built model.

5.3.4 Error cost consideration

There are two types of error in construction firms insolvency prediction. Type I error where an insolvent construction firm is wrongly predicted as healthy, and type II error where a healthy construction firm is wrongly predicted as failing. It is common knowledge that type I error is costlier than type II error. For example, the cost of a firm that takes on expansion or profit spending steps because it thinks it is very solvent/healthy when it is failing is much more than that of a construction firm taking remedial steps because it thinks it is failing when it is solvent/healthy. Also, the cost of awarding contracts to an impending contractor who might fail will typically be much larger than the cost of rejecting a healthy contractor.

To consider error cost in developing a CF-IPM, it is either a study reports overall accuracy and error values for both type of errors, or simply use sensitivity analysis employing the receiver operating characteristic (ROC) curve where the area under the curve (AUC) represents the accuracy of the model. The curve is drawn by plotting type II error against one minus type I error (see section 8.5.3 for more). The curve generalises various compared performances through all achievable cut-off points associated with the error costs and gives some form of cost-benefit analysis for decision-makers (Hosmer, Lemeshow, and Sturdivant 2013; Tserng, Lin *et al.* 2011). With sensitivity analysis, model developers can develop models that will minimise the costlier error by ensuring that a failing construction firm is barely ever predicted as being healthy. Although this increases the confidence of users, it causes an increase in the less costly error

Table 5.1 shows that only 50% of the studies considered error type either directly or through the use of sensitivity analysis; this gives a poor outlook. However, a further breakdown reveals that majority of the recent studies have embraced this criterion in their studies (Figure 5.7) showing a positive trend, especially as most of the primary studies are relatively recent. This trend shows the slow adoption by CF-IPM studies. However, the recent surge in error cost consideration will bring more confidence to CF-IPM users since it ensures that a failing firm is barely ever mistakenly predicted as a healthy one. This means stakeholders of firms predicted as healthy can be double sure the firm is healthy while stakeholders of firms wrongly predicted as failing will take steps that will ensure their firms become even healthier thereby losing almost nothing. This will also ensure bankers or clients never give loan or contract respectively to a failing construction contractor/firm. The only disadvantage in this case, which is less costly compared to other explained options, is the possibility of a healthy construction contractor/firm missing out on loan or contract because it is wrongly predicted as failing. Overall, despite being the better technique, only 26.7% of the studies used sensitivity analysis for error type consideration, representing a poor outlook.

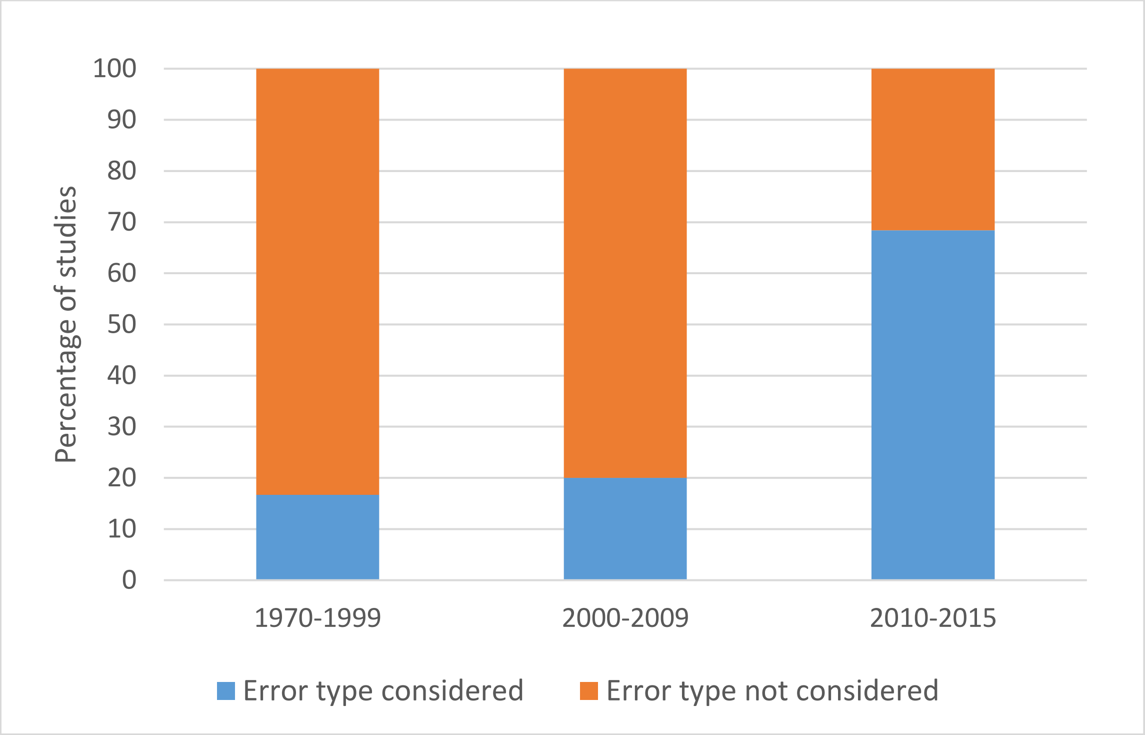


Figure 5.7: Proportion of studies that considered error types over different periods

Since sensitivity analysis is the more sophisticated method of error consideration which gives users more confidence, it is the technique that will be adopted for all the models built in the research to allow for fair comparisons

5.3 Chapter summary

This chapter presented a literature review on the methodical issues in building CF-IPMs, looking at the tools used, characteristics of data or sample used, the model validation process, and the consideration of error cost of the model. While it appears that there is an improvement in relation to some method trends, improvement on other methods appears quite stagnant. The review in this chapter shows that the use of advanced artificial intelligence (AI) tools in building models has been better embraced since 2010 but with the limitation of use to mainly ANN, SVM and DT. There is a need to adopt and compare (models created by) other high performing AI tools like random forest, Adabag Boosting, Extremely Randomized Trees, Naive Bayes, Clustered Support Vector Machines, among others. This will ensure the very best model is selected from the multitude produced from these high performing tools. This is hence the method that will be adopted in this work

The use of skewed data made up of many existing construction firms and few failed construction firms still occurs in the CF-IPM world despite the popular knowledge that, with all tools, it leads to skewed accuracy in favour of the larger group. The argument, used in few of the studies that used skewed data, that applying data balancing techniques like over (under) sampling leads to bias does not relatively hold water. These sampling techniques are well established and there are many cases where the total population cannot be used due to the size. Moreover, the skewed accuracy is even more biased and easily leads to a costlier error of predicting a failing construction firm as a healthy one. The proposed solution will hence use data with equal or almost equal dispersion, employing the undersampling technique with matched samples.

Another unacceptable feature of CF-IPM studies is the poor or non-existent validation technique, where the data used to build a model is used to test it. It is disappointing to have as many more than a third of the primary studies to be involved in this. It is not excessive to say any study that does not validate its CF-IPM with a separate data has not tested it for practical use hence its results should not be accepted. The approach to be used in the proposed solution is thus to set aside a minimum of 20% of the data for testing any built model.

Like the use of AI tools, the consideration of error cost in testing models has been well embraced since 2010. However, despite being the better technique of error type consideration, only 26.7% of the systematically reviewed studies used sensitivity analysis for error type consideration, representing a poor outlook. Since sensitivity analysis is the more sophisticated method of error consideration which gives users more confidence, it is the technique that will be adopted for all the models built in the research to allow for fair comparisons.

The first part of the chapter six contains a review of methodological perspectives of CF-IPM studies with focus on identification of key methodological flaws to be avoided in the research work. The second part contains an explanation of the methodology used in the research work

CHAPTER SIX

6.0 RESEARCH METHODOLOGY

6.1 Chapter introduction

The research methodology used in construction firms insolvency prediction model (CF-IPM) studies has become fixated over time as a result of the copycat approach of the studies. Although the fixation would not be a problem if the methodology were right or optimal, it is not so in this case. The refusal of studies to even look into potential improvement of the methodology of developing CF-IPMs is in itself condemnable. This chapter uses a systematic review method to review the methodological positions of CF-IPM studies and consequently identify areas that can be improved upon, with the improvements implemented for the research.

Section 6.2 is a systematic review of the methodological positions of CF-IPM studies with more focus on paradigm (subsection 6.2.1), ontology and epistemology (subsection 6.2.2). Section 6.3 explains the implication of the narrow methodology CF-IPM studies are fixated on. This is followed by proposed improvements that are adopted in the research in section 6.4. Section 6.5 justifies the unit of analysis of the research. Section 6.6 and 6.7 explain how the qualitative and quantitative aspects of the research were executed respectively, with 6.7.1 and 6.7.2 describing the execution of the survey and company documentation strategies. Section 6.8 is used to summarise the chapter.

6.2 Systematic review of construction firms insolvency prediction model studies with focus on methodological positions

The method used to carry out the systematic review in this section is the same as that presented in section 5.2 which is based on the review in section 3.3. As done in section 5.2, Hall's (1994) study is excluded. In essence, all the 30 systematically reviewed studies in section 5.2 are examined here. To assess the methodological positions of the 30 primary studies (i.e. the systematically reviewed studies), the paradigm or philosophical underpinning, ontology, epistemology and research approaches are appraised. The details to support this review can be found in the ‘Summary of Findings’ table (Table 5.1) in chapter five.

6.2.1 Paradigm of construction firms insolvency prediction model studies

This subsection explores the philosophical underpinning or research paradigms of construction firms insolvency prediction model (CF-IPM) studies. Thomas Kuhn, who popularised the idea of a paradigm, defined *paradigms* in Kuhn (1962) as "universally recognised scientific achievements that, for a time, provide model problems and solutions for a community of researchers". Paradigm, according to relatively recent studies (Guba 1990; Johnson and Onwuegbuzie 2004; Scotland 2012) is a research culture, a set of assumptions, values and belief, which comprise of but is not limited to epistemology, axiology, ontology, methodology, and aesthetic beliefs.

The data collection trend in CF-IPM studies appears to be that of independent observers as most of the primary studies either used only financial ratios, or financial ratios in combination with stock market information (Table 5.1). A few others used stock market information alone while only three studies used some form of non-financial variables (Figure 6.1). All variables are generally used to measure the health of a construction firm, giving the studies a positivist outlook since the positivism paradigm believes that research can mainly be done by observations and measurements (Trochim and Donnelly 2008).

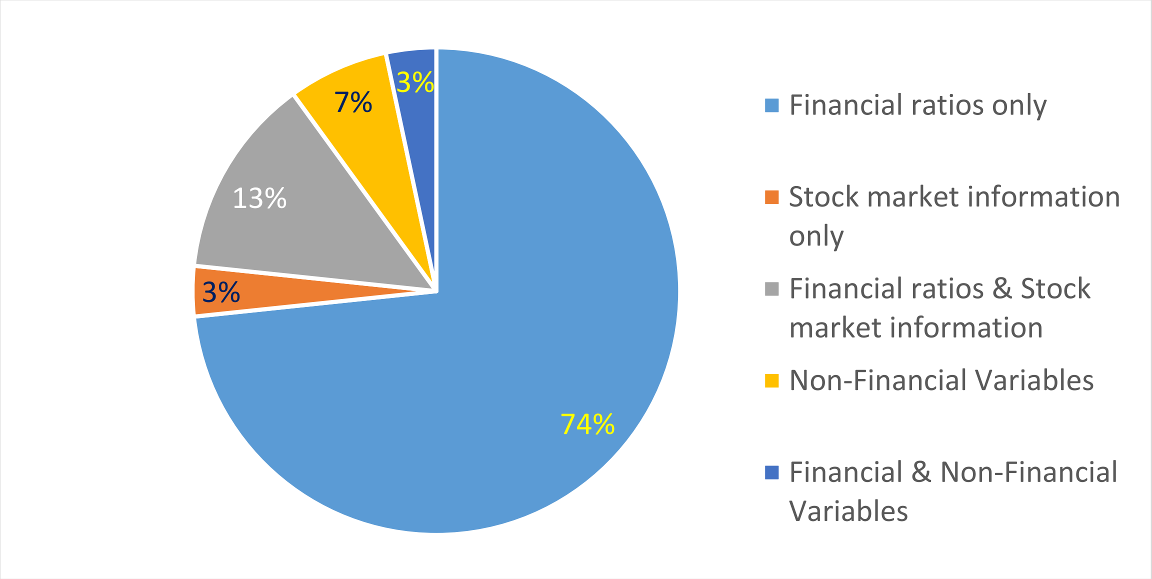


Figure 6.1: Types of variables used in the primary studies

While financial ratios data was collected from some form of financial databases, stock market information was collected from stock exchange organisations e.g. New York Stock

Exchange. These financial variables (financial ratios and stock market information) are common in CF-IPM studies mainly for two reasons:

1) Financial ratios are the variables used by the two successful pioneering studies (i.e. Altman, 1968; Beaver, 1966) that most IPM studies are emulating

2) Financial data are usually readily available from third party or in publicly available company archives and thus makes data collection very easy for a researcher (Laitinen 1992; Dirickx and Van Landeghem " ' 1994).

Of the three primary studies that used non-financial variables **(see Table 5.1),** Horta and Camanho (2013) combined three strategic variables with six financial ratios; Abidali and Harris (1995) built two separate models, one with seven financial ratios and another with 13 managerial variables; while Koksal and Arditi (2004) used a large number of non-financial variables. Horta and Camanho (2013) chose their strategic variables from their previous study. The value for each of the three variables (company main activity, company size and headquarter geographic location) was accessible from company archives and financial databases. Abidali and Harris (1995) and Koksal and Arditi (2004) got the non-financial variables from the literature and used questionnaire to collect the data. Both questionnaire and archival data (from databases) are forms of independent observation which is a positivist approach. A positivist researcher is normally independent (of the subject) as an observer, reduces a phenomenon to simpler measurable factors/elements, explains the elements in terms of how they affect the phenomenon (cause and effect) and usually uses large samples (Burrell and Morgan 2008; Easterby-Smith *et al.* 1991). Positivism seeks to explain and predict what happens in the social world by searching for patterns and relationships (Burrell and Morgan 2008).

In CF-IPM studies, it is the complex failure process (phenomenon) that is reduced to measurable variables, usually financial ratios (simpler elements) measured from databases. The relationship between each variable and the failure process is then explained in the studies and the importance of each variable highlighted, usually through a statistical process, before they are used. Example quotes of where primary studies explained a used variable relationship to failure are as follows:

> The ratio of turnover to net assets is a "*measure of how well a company has used its productive capacity*" (Abidali and Harris, 1995: p.191).

> The activity ratio measures “*how well a company has been using its resources*” (Ng, Wong, and Zhang, 2011: p.601).

Apart from Langford, Iyagba, and Komba (1993) who simply tested an existing model using three firms, and Stroe and Bărbuță-Mișu (2010) that used a sample size of 11 construction firms, the least sample size in the primary studies is 40 construction firms. The mean average sample size of the 29 studies that clearly indicated their sample sizes is 4930. [Note that studies with the largest sample sizes did not use optimally tuned artificial intelligence tools hence might not have executed onerous computing like in Du Jardin's (2010) study where 30 computers were needed for analysing a sample of 500 firms]. Marshall, Cardon, Poddar, and Fontenot (2013) in their comprehensive review of sample sizes in qualitative research, using 83 qualitative studies from top international journals, clearly proved a sample size of 30 to be high and that saturation is normally reached before reaching this number (i.e. 30). Figure 6.2 shows the sample size ranges used in the primary studies. It is clear from the figure that majority of the studies used a large sample size, well beyond the 30 limit in qualitative studies, which clearly depicts them as quantitative studies. In fact, more than 50% of the studies used more than 100 sample firms. Using large samples and quantitative methods, as noted earlier, is a feature of positivism.

In positivism, research is “seen as the way to get at truth, to understand the world well enough so that we might predict and control it” (Trochim and Donnelly, 2008: p.18). This is exactly the aim of most CF-IPM studies. In the studies, an attempt is made to understand failure of construction firms and to identify failure indicators; then there is an effort to predict potential failure in order to aid control of the situation by owners taking mitigating steps or financiers making decisions on loans. The aim of CF-IPM studies thus in itself, to an extent, lend them to positivism.

According to Burrell and Morgan (2008), the functionalist/positivist is always seeking to find implementable solutions to real problems and is more concerned with controlling social affairs. This is well in line with the aim of CF-IPM studies which try to provide IPM as a solution to the real problem of either high rate of construction firms failure or to the problem of identifying healthy companies for loans or contract. CF-IPM studies have used mainly quantitative data, usually in the form of financial ratios, which is a common feature of positivism (Phillips and Burbules 2000; Mukherji and Albon 2010). Further, positivists tend to use statistical analysis so as to aid generalisation (Alvesson and Sköldberg 2000;

Mukherji and Albon 2010); this is typical of CF-IPM studies since the model (i.e. the CF-IPM) is built using a statistical method. Note that artificial intelligence (AI) tools are advanced statistical/mathematical methods.

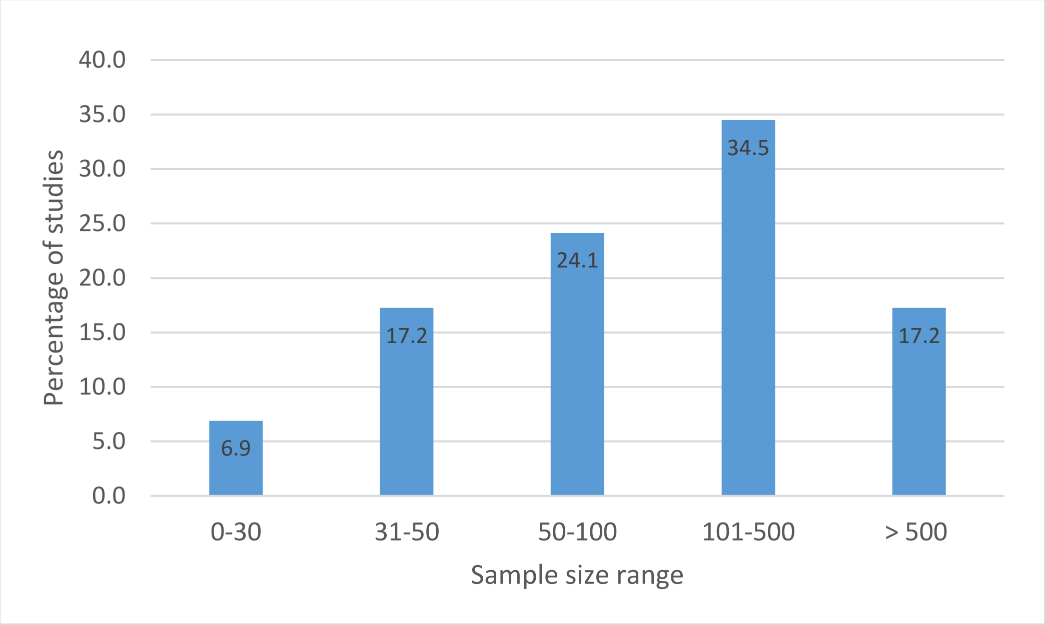


Figure 6.2: Sample size ranges used in the primary studies. Many of the primary studies used a large sample size.

From all the evidence given in this subsection, it appears that the positivism/functionalism paradigm is predominant in the CF-IPM literature. This is well understandable since prediction, the main aim of the studies, is a main feature of positivism. Although critical realism also supports quantitative data and analysis, and possess some features similar to those of positivism, it is not used mainly for prediction. A critical realist is also not an independent observer, i.e. an objectivist, as is with CF-IPM researchers. A brief look at the ontology and epistemology of the reviewed studies can shed more light on this area of discussion.

6.2.2 Ontology, epistemology and research approaches of construction firms insolvency prediction model studies

Ontology deals with the assumption researchers have on how knowledge exists (Burrell and Morgan 2008). It is defined by Blaikie (2007) as the science or study of being and it deals with the nature of reality. Epistemology deals with how to learn that reality/knowledge

(Burrell and Morgan 2008). The realist ontology and objective epistemology are features of positivism (Kolakowski 1972; Burrell and Morgan 2008; Easterby-Smith *et al.* 1991) and are the adopted forms in CF-IPM studies. Realism "assumes that social and natural reality exist independently of our cognitive structure: an extra-mental reality exists whether or not human beings can actually gain cognitive access to it" (P. Johnson and Duberley, 2000: p.67). The realism ontology is in itself quite embedded in the nature of CF-IPM enquiries since the statistics of mass failure of construction firms is repeatedly available in many financial and government reports. The failure is real whether or not human beings can access, assess, prevent or hasten it, or whether human beings know about it at all or not. This is pretty much the opposite of idealism ontology which "assumes that what we take to be external social and natural reality is merely a creation of our consciousness and cognitions" (Johnson and Duberley 2000).

Epistemology wise, objectivism is the widely used option in CF-IPM studies. Objectivism accepts that reality and its meaning exists independent of any awareness or recognition and can be learned (Crotty, 1998); it focuses on the object with absolutely no regards for the subjects (Burrell and Morgan 2008). CF-IPM studies are directly concerned with only the object i.e. the construction firms. In the primary studies reviewed, developing the CF-IPMs was done in virtually all cases with absolutely no contact with the subject i.e. any representative of the sample construction firms (e.g. owner, employee, firm's lawyer, among others). The information used to develop the CF-IPMs were mainly in the form of financial variables gotten from financial databases and stock exchange organisations (e.g. New York Stock Exchange), independent of the subjects of sample construction firms. In the very rare cases where non-financial variables are used, questionnaires, which are also objective, are used to get the variables. The exclusive use of the objective approach in CF-IPM studies has however been an area where improvement can be made since it has always been an area of contention between experts, plus the construction industry is quite dynamic.

The use of information from databases and questionnaire implies the use of archival and survey research strategies respectively. The research approach of the primary studies can also be concluded to be deductive, another feature of positivism (Easterby-Smith, Thorpe and Jackson, 2008). A deductive approach is used when there is plenty of literature on the research area and one of the existing theories in the literature is to be tested (Easterby-Smith, Thorpe and Jackson, 2008; Holloway, 1997; Robson, 2011). Nearly all the primary studies initially collected financial variables from existing literature and then followed in the footsteps of previous studies (e.g. Altman, 1968; Edmister, 1972; Ohlson, 1980; Zavgren, 1983; among others) to build their models. They always tested the theory that a certain selection of variables and a/some statistical tool(s) (including AI tools and option-based models) can be used together to build a high performing model for predicting failure of construction firms.

6.3 Implication of the narrow methodological positions in CF-IPM studies

The restricted use of positivism in CF-IPM studies has led to the continuous, exclusive use of the objective epistemology, through the use of multivariate analysis of financial ratios. Unfortunately, this singular dimension approach does not fully represent the insolvency situation of construction firms as highlighted in various studies; and due to the dynamism of the construction industry.

On facts highlighted in various studies, a countless number of non-financial indications of insolvency, such as management mistakes, do come up a lot earlier than financial distress (Abidali and Harris 1995). Financial distress only tends to be noticeable when the failure process is almost complete, around the last two years of failure according to Abidali and Harris (1995). The truth is that it is adverse managerial actions and other social factors that lead to poor financial standings and in turn cause insolvency. Accordingly, many management experts have reiterated that financial variables alone are insufficient for the early depiction of disastrous factors like shambolic management, acquisition of a failing construction firm, economic decline, among others (Argenti, 1980). Many construction firms insolvency researchers (e.g. Abidali and Harris, 1995; Arditi, Koksal, and Kale, 2000; Kale and Arditi, 1999; Kangari, 1988; among others) have also stressed that financial ratio models are not enough to predict insolvency of construction firms until they are used with other economic, managerial and social factors. Further, the tendency of accountants to amend important financial ratios, known as window dressing or creative accounting, reduces the reliability of financial ratios as factors for predicting insolvency (Arditi *et al.*, 2000; Argenti, 1976; Balcaen and Ooghe, 2006; Rosner, 2003; among others) [Please see chapter three for more].

In addition, in many countries (including the UK and France among others) only certain firms that meet some specific criteria like a pre-set minimum asset size, number of employees, among others, are required by law to produce financial statements periodically. Hence micro, small and medium construction firms which make up an overriding majority of the construction industry might not have periodic financial statements (Balcaen and Ooghe 2006). The statistics of the UK construction industry are clear: the industry boasted over 950,000 small and medium enterprise (SME) in 2015; the industry represents circa 20% of the UK private sector SMEs, making it the sector with the highest percentage of SME firms (Department for Business Innovation and Skills, 2015); over 96% of UK construction firms as of 2001 are small or micro firms (Jaunzens, 2001); and 86% of employees in the sector work in small construction firms (Jaunzens, 2001). A lot of SMEs do not have a good accounting system and hence only produce/submit incomplete and inadequate financial statements; this automatically nullifies the possibility of their involvement in CF-IPM studies since incomplete financial statements are normally discarded of in IPM studies (Tucker 1996; Balcaen and Ooghe 2006). This means most CF-IPM studies build models that are not applicable to small construction firms despite the fact that they (i.e. micro firms and SMEs) are well known to make up a larger percentage of the failing firms.

Another problem is that some SMEs do outsource their account management to independent accounting firms with the sole intention of ensuring periodic production of their financial statement in order to satisfy legal/government requirements. This sometimes leads to misrepresented financial statements based solely on the amount of information provided to the accounting firm by the construction firm. In a similar fashion, some SME firms simply produce poor and inaccurate statements themselves simply to fulfil the legal requirements. Any CF-IPM developed from such statements will have limited practical usefulness

On the dynamism of the construction industry, the dynamic nature of the industry with constantly changing trends (Chang, 2001; Chen, 2009; R. Navon, 2007; Ronie Navon, 2005; Odusami, Iyagba, and Omirin, 2003; Razak Bin Ibrahim, Roy, Ahmed, and Imtiaz, 2010; among others) means the main causes of failure of construction firms will vary from time to time. This implies that key players like owners, directors, managers, among others (i.e. subjects), will have to be spoken to in order to understand key reasons behind failure of construction firms at different times. Ultimately, leaving out the subjects appears not to be a wise choice if a valid CF-IPM is to be built.

The need to involve social factors, which can mainly be considered through a subjective approach, and the need to talk to subjects to understand the timely dynamics of the construction industry, both call strongly for the adoption of the subjective epistemology in CF-IPM studies. Remenyi (1998: p.35) stressed the importance of investigating “the details of the situation to understand the reality or perhaps a reality working behind them”. This is only achievable subjectively since subjectivism emphasises on seeking explanation to understand a social phenomenon (Burrell and Morgan 2008). A very good understanding of the failure process of construction firms by a CF-IPM developer will definitely contribute to an improved model.

6.4 Proposed improvements adopted for the research’s methodology

Concerning paradigms, having reviewed numerous CF-IPM studies, the paradigm proposed for the research is pragmatism. Pragmatism argues that the main determinant of the methodology to be used in research should be the research questions rather than strictly following a particular paradigm because of a sociological belief, or so as to copy past studies as done in CF-IPM studies (Johnson and Onwuegbuzie 2004). Pragmatists are more concerned with the practical consequences of the research findings and as such believe that one standpoint can never be suitable for answering all types of research questions and there may be multiple realities (Dewey 1920; Murphy and Rorty 1990; James 1995). This is the maximalist view noted by Callon (2006) and Johnson and Onwuegbuzie (2004) which argues that nothing in a research phenomenon can escape pragmatics. Pragmatists neither agree with positivists in that demands of a research cannot be fully satisfied by a theory (falsify-ability, objectivity, among others), nor with interpretivists in that demands of a research can be satisfied (at least partly) by almost any theory (Powell, 2001). This is in similarity to the Actor-network theory (ANT) which "privileges neither natural (realism) nor cultural (social constructivism) accounts of scientific production, asserting instead that science is a process of heterogeneous engineering in which the social, technical, conceptual, and textual are puzzled together (or juxtaposed) and transformed (or translated)" (Ritzer, 2004: p.1). Pragmatism thus allows the use of any, or a mix of multiple methods, approaches, choices, tools, among others, as long as they will help to answer the research questions properly (Johnson and Onwuegbuzie 2004). It allows the researcher to “study what interests you and is of value to you, study in the different ways in which you deem appropriate, and use the results in ways that can bring about positive consequences within your value system" (Tashakkori and Teddlie, 1998: p.30).

The relative rigidity of other paradigms as to the methodological positions that fit a research can limit steps needed to be taken to complete quality research.This is confirmed by Saunders, Lewis, and Thornhill, (2009: p.109) that "the practical reality is that a particular research question rarely falls neatly into only one philosophical domain". Further, a good CF-IPM study should focus on failure of construction firms (a problem) experienced by construction firms owners (people) and the effect of developing a CF-IPM which will allow timely intervention that can prevent potential failure (consequence of inquiry). Such focus is synonymous with pragmatism which "emphasises the practical problems experienced by people, the research questions posited, and the consequences of inquiry" (Giacobbi, Poczwardowski, and Hager, 2005: p.18).

The realist ontology used for CF-IPM studies is very appropriate and is consequently used here. There is only one reality, and it is that 'construction firms do fail and failing construction firms have certain similar attributes'. Finding the most effective attributes to develop a CF-IPM is what is tricky. This is one of the reasons there are many CF-IPM studies, each trying to prove certain attributes are more effective than others.

Although the objective epistemological stance is suitable for developing a CF-IPM, a combined subjective and objective approach in a facilitation manner is proposed and used here. While the objective approach will aid the use of existing factors and variables, the subjective approach can be used to identify temporal factors and variables that can be used to develop a timely and robust CF-IPM; this would have taken the dynamism of the construction industry into consideration. The subjective approach can also help identify important social and managerial factors that contribute to insolvency of construction firms. This has long been advocated by many construction management (CM) authors (e.g. Dainty, 2008; Seymour, Crook, and Rooke, 1997), who queried the focus on objects when at the centre of most CM research is people (subjects), justifying the need for greater emphasis on qualitative enquiry. Management level staff and owners of failed and existing construction firms can use their practical experience to contribute vital information regarding factors that affect insolvency and survival of construction firms hence they need to, and will be engaged in the research.

Since both the objective and subjective epistemology are vital for a valid CF-IPM, the integration of quantitative and qualitative research approach is proposed and used in the research. This is in line with the much advocated methodological pluralism in CM (Seymour *et al.* 1997; Mingers and Gill 1997) which combines methodologies from varying paradigms to provide richer insights into relationships and their interconnectivity (between factors and firm failure in this case); this is the best approach to solving research problems (Mingers and Gill 1997). The use of the dual epistemology approach is only possible with mixed methods which in itself is a feature of pragmatism (Hoshmand 2003; Johnson and Onwuegbuzie 2004). Mixed method can combine the strengths of different methods to provide a more robust approach to answering a research question and avoid preconceived biases (Sechrest and Sidani 1995) which might exist in past studies that are being copied. For example, many CF-IPM studies wrongly ended up using unequal data dispersion simply by copying methodologies of past studies

In this vein, the proposed methodology to be used in the research agrees with the popular Seymour and Rooke's (1995) work which clearly argued that different researches require different methods and no method should be ruled out a priori. However, it does not support their opposition to the multi-paradigm (see Rooke, Seymour, and Crook, 1997) approach which pragmatism allows if it is what will bring about a valid methodology to answer the research question in focus. In fact, such opposition is tantamount to nullifying some methods a priori since selecting a particular paradigm readily nullifies some methods; an act Rooke *et al.* (1997) themselves preach against. An improved research methodology framework for developing a CF-IPM which is used in the research is given in Figure 6.3.

The subjective epistemology aspect of the work is executed with the multiple case study strategy. Case study is defined by Mitchell (1983: p. 192) as a “detailed examination of an event (or series of related events) which the analyst believes exhibits (or exhibit) the operation of some identified general theoretical principles”. Yin (1994: p. 13) defined a case study as “an empirical inquiry that investigates a contemporary phenomenon within its real-life context, especially when the boundaries between phenomenon and context are not clearly evident. Yin (1994) went ahead to explain that case studies usually require more than one source of evidence. The case study will be executed using interviews

The temporal factors obtained (because the construction industry is dynamic) will be analysed to identify befitting measuring variables which can then be measured with a survey research strategy. This process of identifying new factors and discovering patterns from the field culminates in an inductive research approach (Easterby-Smith, Thorpe and Jackson, 2008).

The objective epistemology aspect, which is the quantitative study, will be executed using survey and archival research strategies. Survey will be executed with a Likert scale questionnaire. Archival research strategy, which involves collecting financial data from companies' archives, financial databases or stock market is the norm in CF-IPM. The term 'archival' in this strategy does not directly mean 'old' in any way as pointed out by Bryman (1989) hence using recent financial statements also fall under this category.

This proposed strategy culminates in facilitation which involves the "use of one data collection method or research strategy to aid research using another data collection method or research strategy within a study" (M. Saunders and Paul, 2013: p.154). In the research, the unstructured interview data collection method aids the questionnaire data collection method. The proposed strategy also shows the intended mixed method approach (qualitative and quantitative data collection and analysis). The mixed method approach is very good since it ensures an all-round effectiveness of research (Creswell and Plano Clark 2011) and is well in line with the proposed pragmatism philosophical stance (Giacobbi, Poczwardowski and Hager, 2005) in the research. The methodological positions taken in the research are summarised in Table 6.1

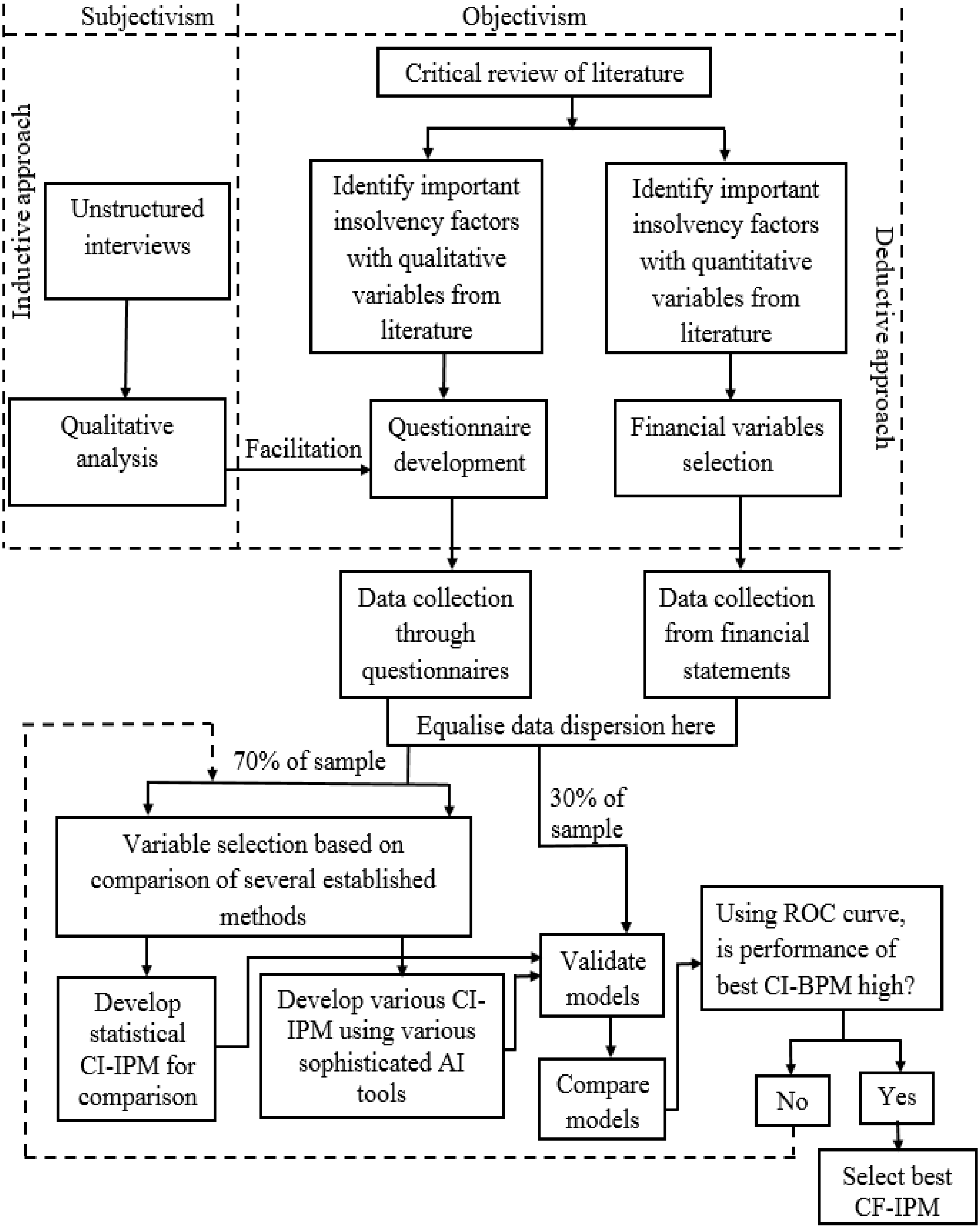

Subjectivism
Objectivism
Inductive approach
Deductive approach
Critical review of literature
Unstructured interviews
Identify important insolvency factors with qualitative variables from literature
Identify important insolvency factors with quantitative variables from literature
Qualitative analysis
Facilitation
Questionnaire development
Financial variables selection
Data collection through questionnaires
Data collection from financial statements
Equalise data dispersion here
70% of sample
30% of sample
Variable selection based on comparison of several established methods
Develop statistical CI-IPM for comparison
Develop various CI-IPM using various sophisticated AI tools
Validate models
Compare models
Using ROC curve, is performance of best CI-BPM high?
No
Yes
Select best CF-IPM

Figure 6.3: An improved research methodology framework used for developing CF-IPM in the research (AI: artificial intelligence, ROC: receiver operating characteristics)

Table 6.1: Critical Choices in Research Designs

Areas of Research Options	**Options available**	**Appropriately selected option**
Research Philosophy or Paradigm	1. Positivism 2. Interpretivism 3. Critical realism 4. Direct realism 5. Pragmatism	Pragmatism
Ontology	1. Realism 2. Idealism	1. Idealism and 2. Realism
Epistemology	1. Objectivism 2. Subjectivism 3. Constructivism	1. Subjectivism and 2. Objectivism
Research Approach	1. Deduction 2. Induction 3. Abduction 4. Retroduction	1. Induction and 2. Deduction
Research Strategy	1. Experiment 2. Survey 3. Case Study 4. Action Research 5. Grounded Theory 6. Ethnography 7. Archival Research	1. Case Study, 2. Survey and 3. Archival Research
Type of Case Design/Studies	1. Single-Case Designs 2. Multiple-Case Designs	1. Multiple-Case Designs
Research Methods/Choices	1. MonoMethod 2. Multi-Method 3. Mixed Method	1. Mixed Method
Reasons for Using Mixed Method	1. Triangulation 2. Facilitation 3. Complementarity 4. Aid Interpretation and more	1. Facilitation
Data Collection Methods	1. Direct Observation 2. Interviews 3. Focus Group Discussion 4. Questionnaires 5. Company Documentation 6. Reporting	1. Interviews (unstructured), 2. Questionnaires and 3. Company Documentation

Areas of Research Options	Options available	Appropriately selected option
Data Analysis Techniques	Many	1. Thematic analysis 2. Reliability analysis 3. Factor analysis among others

6.4.1 *Justification for using interviews*

The case studies in the research was executed with the storytelling strategy using unstructured interviews. The unstructured interviews for case studies was used to get answers to questions like: what are the common factors that lead to insolvency and how do these factors affect insolvency? This reason goes down well with case study's superb capability of obtaining answers to the 'what?' and 'why?' questions (Saunders, Lewis and Thornhill, 2000).

Since the intention was to get a limitless an in-depth examination of how construction firms fail based on the experience of respondents, there were no pre-determined or pro-ordered set of questions; the unstructured interview is meant for this sort of scenarios (i.e. no pre-set questions) hence the reason it was adopted. "The unstructured interviews take the form of free-flowing conversation" and are known for the advantage of not limiting respondent views (Latham and Finnegan 1993; p. 42). The flexibility of the method (i.e. unstructured interview) allows an interviewer to ask further questions on any part of the answer of the interviewee, thereby giving the research the opportunity to destroy ambiguity at the data collection stage. Further, it was believed that the informal setting which comes with unstructured interview would help put the respondents in a relaxed and comfortable mode, thereby letting them freely provide as much information as possible.

One issue common with the investigation of construction firms failures (not just CF-IPM studies) is that when interviews are used, the respondents are usually asked for their views when some of them cannot judge best what some key problems were as they have repeatedly failed with subsequently established firms. Those unidentified key problems are referred to as the deeper truths which are unattainable with direct observation; a viewpoint rejected by positivism and empiricism but well accepted by structuralism, hermeneutics and psychoanalysis (Gabriel and Griffiths 2004). On using the subjective approach (i.e.

interviews) to search for the deeper truth, it is usually onerous to detach the more or less important insolvency criteria by respondents in research. Ordinarily, the owner, manager, employee, among others, of a failed firm is more tilted to blaming other stakeholders although, such blames are sometimes true. The research thus elucidated the complex process of failure of construction firms by analysing the 'stories' of owners, directors, managers and (or) employees of failed and existing construction firms as was deemed fit. This was done by listening to their accounts of the life of the construction firm from its establishment (or stage of involvement) to insolvency (where applicable) or time of interview. By using the storytelling method, any form of prior assumptions about the criteria that lead to insolvency was prevented and a chance to conduct a narrative analysis of the stories to identify what events, actions, or occasions contributed to (in)solvency was created. Storytelling can be of unstructured interview as was in the research, or semi-structured in other cases (Gabriel and Griffiths 2004). The adopted process ensured that the first objective of the research was met (see section 1.7)

6.4.1.1 Advantages and disadvantages of interview

Interviews have a number of pros and cons. The major advantage of unstructured interview is that it allows respondents to explain things to the fullest and allows ambiguity to be cleared immediately (Merton and Kendall 1946). It allows issues to be investigate in depth and can be a source of lead to other respondents (Ryan, Coughlan and Cronin, 2009). These advantages allowed a comprehensive understanding of construction firms insolvency to achieved in the qualitative aspect of my research. Interviews can also be relatively inexpensive as the require a relatively low number of respondents; saturation is normally reached after only 12 interviews (Guest, Bunce and Johnson, 2006)

On the negative side, unstructured interviews can yield a lot of unimportant information since the method restricts the intervention of interviewee while the respondent is talking (Hycner, 1985). The method also encourages the respondent to talk as much as possible and can be time consuming in the case of a highly willing talkative respondent (Roberta and Cowton 2000). The subjective nature also means that respondents can share personal views which they think affect the situation under examination but do not, thereby giving inaccurate information. This was guarded against by using the story telling method. Finally, interviews can be difficult to analyse as was the case in the research.

6.4.2 *Justification for using questionnaire*

The survey strategy was executed with the questionnaire method. To stimulate responses from target respondents, the Likert scale questionnaires with closed-ended question was used since it is quite easy to deal with by respondents (Van Laerhoven, van der Zaag-Loonen and Derkx, 2004). The use of Likert scale questionnaire was important in the research because it represented a way of allowing respondents to rate the extent to which each qualitative variable (factor), identified from the unstructured interviews, applied to their construction firm. This was vital because the ratings put some form of numbers on the qualitative variables. The use of these numbers was the only way the qualitative variables could have been used together with the quantitative variables, as input variables during the development of the CF-IPM; this is a key objective of the research (see objective number three in section 1.7).

6.4.2.1 Advantages and disadvantages of questionnaire

The questionnaire method has an advantage of ensuring that all respondents are exposed to exactly the same questions and are given exactly the same options to pick from (Foddy, 1993). This helps to avoid any potential bias in the questioning method. It also eliminates the case of a respondent forgetting to give some relevant answers. A major advantage of Likert scale questionnaire is that it helps to provide structured data which can be easy to analyse (Smith and Hakel 1979). This was very helpful for the initial analysis in the research, when reliability and factor analysis were carried out (see subsections 7.3.1 and 7.3.4). Also, responses can be gathered from a large number of respondents as done in the research, thereby helping to improve reliability

On the flip side, questionnaires can be expensive to distribute and collect if it is not done online using free mediums like Google forms (Wright, 2005). Most of the questionnaires in the research had to be posted, using prepaid return envelopes, to the available addresses on databases. This was done for two reasons: 1) insolvent construction firms had dormant email addresses on databases so questionnaires were sent to the current work addresses of former owner/directors, 2) Many MSM firms did not have an email address at all and response rate from the big firms' respondents through email was critically poor. In addition, with questionnaires, respondents cannot make clarifications on questions they do not understand, and their expressible response/views are limited (Reja *et al.*, 2003). Finally, Likert scale questionnaire provides discrete variables which could be a pro or con. It turned out to be a con here as the discrete data proved to be tricky for the AI tools to handle.

6.5 Unit of analysis

Unit of analysis, according to Tainton (1990: p.5), "is the entity on which there are data and which will be subjected to analysis." This idea is used to define the main entity which is analysed in a study (Trochim and Donnelly 2008). The unit of analysis is therefore absolutely dependent on the design of the study. Although they are usually the same, the unit of analysis in a study might be different from the unit of observation which "is the entity on which the original measurements are made" (Tainton 1990: p.5). The following are the categories of unit of analysis according to (Bless, Higson-Smith and Kagee, 2006):

- *Individuals:* The case where the research studies and analyses a set of individuals that belong to a particular group such as young girls, carpenters, white Muslims, among others. Each individual is a unit. This category is the most popular unit of analysis

- *Groups*: This involves studying different groups and probably comparing the groups. In this case, each group, and not the individual members of the group, represents a unit

- *Organisations*: Organisations are a type of group that is commonly used as unit of analysis in social science research, each organisation in the study representing a unit. Some organisations can be compared based on their profits, proportion of employees of certain background, policy effectiveness, corporate social responsibility, and so forth

- *Social artefact*: These are "products of social beings and can be anything from poems and letters to automobiles and farming implements. A systematic analysis of such artefacts may provide valuable information about the individuals and groups that created or use them" (Bless *et al.*, 2006: p. 73).

- *Period of time*: This involves analysing how something has changed over time.

It is clear from above that unit of analysis and unit of observation are well related to the samples that will be used in a study. The research intends to build a robust IPM for construction businesses to detect potential failure very early. To do this, data about construction businesses have to be taken, carefully studied and analysed hence both the unit of analysis and the unit of observation for the research are construction firms i.e. *organisations as unit of analysis*. The sample construction firms used in the research are the ones whose area of operation are mainly in the UK.

6.6 Qualitative study

6.6.1 Sampling for Interviews

The criteria used to select target participants/respondents were that

1) They are, or were, owners/directors of large, medium, small or micro construction firms. The construction firm could be existing or could have failed
2) They were in the aforementioned position for at least one year.

These are the people in charge of the daily affairs of the firm hence they have a good amount of knowledge/information about the company. Although the preferred minimum number of years of experience was three, the difficulty in getting respondents prompted a change to one year minimum in order to increase the pool to choose from. However, concerted effort was made to get relatively more respondents with minimum of three years' experience and this was partially successful as is evident in tables 6.3 and 6.5.

The database used for sampling in the research is FAME (Forecasting Analysis and Modelling Environment) Bureau Van Dijk UK financial database which contains the details of most of the firms in the United Kingdom. Details in the database include firms' general details like trading address, website, email address, trade, year of establishment, among others. The database also contains contact details of owners, directors, accountants and secretaries of firms.

For existing firms, only random sampling was used. The method was combined with the convenience and snowballing methods for failed firms. The contact for owners/directors of existing firms were gotten by searching FAME the fame database. The search for large construction firms was done separately from that of micro, small and medium (MSM) construction firms. The number of employees option on the FAME database website was used to separate the searches for the construction firm sizes (please see section 1.11 in chapter one for firm size definition by number of employees according to the European Union). The search criteria and the selected options for existing MSM and large construction firms are displayed in Table 6.2.

Table 6.2: Search criteria and selected options for existing MSM and large construction firms

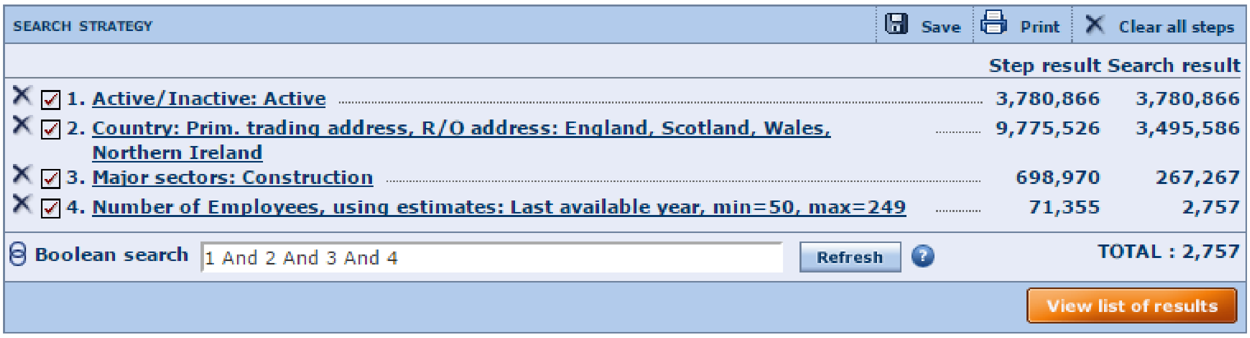
Search criterion	**Selected option(s)**
Active/Inactive	Active (for existing firms)
Major sectors	Construction
Country prime trading address	England Scotland Wales Northern Ireland
Number of employees, using estimates	[min = 1 and max = 249] for micro, small and medium firms search [min = 250 and max = (no input value)] for large firms search

The searches returned over 230,000 MSM construction firms and over 650 large construction firm. For large construction firms, every 10th firm was then selected until 50 firms were selected. For the MSM firms, the search returned, the results were arranged according to number of employees in the descending order so that the first set of firms on display had 249 employees (i.e. medium sized firms). Every 50th firm was then selected until 20 firms were selected (some of which had below 249 employees but not below 50). The 'next' button was then clicked until firms with a maximum number of 49 were displayed (i.e. small firms). Again, every 50th firm was then selected until 20 firms were selected. For micro firms with a maximum of nine employees, every 1000th firm was selected until 50 firms were selected. This approach was used because the search breakdown gave 2757 medium firms, 8962 small firms and over 226,233 micro firms. Where the firm to be selected did not fall under the UK Standard industrial classification of economic activities (SIC) 2007 listed in section 1.11 (scope and limitation) of chapter one, the next one that did so was selected. All the contacts selected were served with an interview request. Table 6.3 presents the demographics of the respondents that agreed to participate. A picture of the summary result of the search for medium (50-249 employees) existing construction firms is given in Figure 6.4 *The process was monitored to ensure the number of respondents for existing construction firms were not a lot more than those for failed construction firms, and vice versa*.

Table 6.3: Demographics of the respondents for existing firms

No. of respondents for large construction firms	**No. of respondents for MSM construction firms**	**No. of years of ownership/ directorship experience with the firm in question**
1 (snowball)	3	1-2
1	2	3-5
0	3	6-10
2	1	11 - 20
0	0	21 and above
Total = 4	*Total =9*	

Note: All respondents of existing firms, as given in this table, were selected based on random sampling as mentioned earlier in this section



	Step result	Search result
☑ 1. Active/Inactive: Active	3,780,866	3,780,866
☑ 2. Country: Prim. trading address, R/O address: England, Scotland, Wales, Northern Ireland	9,775,526	3,495,586
☑ 3. Major sectors: Construction	698,970	267,267
☑ 4. Number of Employees, using estimates: Last available year, min=50, max=249	71,355	2,757

Boolean search: 1 And 2 And 3 And 4

TOTAL : 2,757

Figure 6.4: A picture of the summary result of the search for medium (50-249 employees) existing construction firms

For failed or insolvent firms, random sampling was combined with the convenience and snowballing methods. First, FAME financial database was used to identify directors of construction firms that failed between the years 2009 and 2016, and subsequently to identify existing firms where those directors currently work. This is easy in FAME as it displays the work history, including current positions, of any director whose name is left-clicked on the computer mouse. Most of the directors in the case of MSM construction firms unsurprisingly turned out to be the owner of the firms. The search criteria and the selected options are displayed in Table 6.4.

The searches returned over 159,000 MSM construction firms and over 146 large construction firm. In the random sampling method, for large construction firms, every 2nd firm was then selected until 50 firms were selected. For the MSM firms, the firms were selected the same way as was done for existing firms. The search breakdown gave 1000 medium firms, 3851 small firms and over 154,613 micro firms. For the convenience sampling, the author of this thesis used his position as a part-time college lecturer on construction apprentice programmes. The apprentices were persuaded to talk to colleagues and bosses at work in order to identify those that have worked in, managed or owned a now defunct micro, small and (or) medium construction firms. Some apprentices were, by themselves, suitable respondents as they once owned firms and most agreed to respond positively to the request of talking to colleagues and bosses. Convenience sampling method has been used in some construction studies before (e.g. Li, Akintoye, Edwards, and Hardcastle, 2005; Oyedele, 2013). This sampling method became necessary because of the inherent difficulty in finding stakeholders of insolvent construction firms.

Table 6.4: Search criteria and selected options for failed MSM and large construction firms

Search criterion	Selected option(s)
Active/Inactive	Active (for existing firms) Dissolved (for failed firms) Liquidated (for failed firms)
Major sectors	Construction
*Date of liquidation/dissolution**	On or after 01/01/2009 Up to and including 31/08/2016
Country prime trading address	England Scotland Wales Northern Ireland
Number of employees, using estimates	[min = 1 and max = 249] for micro, small and medium firms search [min = 250 and max = (no input value)] for large firms search

The snowballing sampling was also employed for both size categories of firms in that interviewees were requested to supply contacts sample targets like them if they did not mind. Since insolvent firms are virtually impossible to trace because of their non-functioning-anymore contacts (Everett and Watson 1998; Stokes and Blackburn 2002; Harada 2007), the interviews/stories from the research supplied a unique resource. All of the contacts gotten were served with an interview request. Table 6.5 presents the demographics of the respondents and the firms. The convenience (using position as a lecturer in a college) and snowballing sampling methods yielded 42.86% of the respondents, indicating that the random sampling from FAME search is less effective for failed construction firms

Table 6.5: Demographics of the respondents for failed firms

No. of respondents for Big construction firms	**No. of respondents for MSM construction firms**	**No. of years of ownership/ directorship experience with the firm in question**	**No. of owner respondents that currently own another firm***
1 (S)	2 (R and C)	1-2	1
1 (S)	2 (R)	3-5	2
0	3 (R)	6-10	0
1 (R)	1 (C)	11 - 20	1
1 (S)	2 (R and C)	21 and above	0
Total = 4	*Total = 10*		*Total = 4*

** for MSM firms only*

Note: letters in bracket represent the sampling method used to select/recruit the respondent.
C: Convenience R: Random S: Snowball

6.6.2 Pilot interviews for qualitative study

A total of five construction firms were used for the pilot study. This included two large construction firms (one existing and one failed), and three MSM construction firms (one existing and two failed). All the respondents were people I knew one way or another as they were recruited through the snowballing sampling method explained in subsection 6.6.1. The respondents were made aware that the study was a pilot one. The respondents were simply asked: ‘please tell me the story of (name of the construction firm in question) from when you have known it till failure (or till now in the case of existing firms). Please be free as much as possible and remember that total confidentiality is guaranteed. Please note that no information is irrelevant to me and no stories are unnecessary, as long as they relate to the (name of the construction firm in question) construction firm.

Every other question was generated from the responses given. For example, when a respondent of a failed MSM firm lamented about high immigration levels, I asked him to please expatiate on how this affected the (name of the construction firm in question) construction firm.

The main feedback from the pilot study was that the words ‘insolvency’, ‘failure’, ‘bust’ or their synonyms should be avoided by the interviewee since some respondents might still feel bad about their firm’s insolvency. The respondent unanimously agreed to the validity of the questions.

6.6.3 Execution of the story telling interviews

As explained in subsection 6.6.2, the questions used in the interviews were designed such that they were not restricting to avoid pre-determined responses and to evoke stories about how the firm’s failure (or survival) came about. Although it was referred to as being in its infancy stage in 2004 (Gabriel and Griffiths 2004), the storytelling method is now a widely accepted and used method (see for example Bouwen and Steyaert, 1997; Hill and McGowan, 1999; Marcella and Illingworth, 2012; Rae, 2000 among others). In fact, Denning (2005) emphasised that research that does not value storytelling as a way of understanding firm performance cannot give a complete account of that firm.

Storytelling or narratives are taken to be especially valuable and appropriate when researching sensitive topics such as insolvency of firms (Marcella and Illingworth 2012). Insolvency can be a bad experience for some owners which they do not want to recall or discuss. Extra effort was thus made to make the questions as non-judgemental as possible.

More time was spent with respondents that delivered more or longer stories; this what is required when the story topics (i.e. construction firms in this case), as against the storytellers, are the unit of analysis (Gabriel and Griffiths 2004) as is the case in the research. Incidents that related to insolvency or firm problems were explored further after the stories by seeking elicit accounts of the incidents through direct or indirect tactic; this is appropriate for the storytelling method according to Gabriel and Griffiths (2004).

In the case of insolvent firms, the stories elicited from the respondents of MSM construction firms can be categorised as tragic considering the four categories of stories (comic, epic, tragic and romantic) presented by Gabriel and Griffiths (2004). This is not too surprising as many owners of insolvent MSM construction firms were not happy about the insolvency. Some stories, however, sounded epic, or a combination of tragedy and epic, as the respondents tried more to show how they made mistakes and learned from them and then defiantly started (or are willing to start) another firm which is now (will be) a success. For big firms, most of the respondents practically blamed the members of the senior management team and barely found their contribution to the failure of the firms.

6.7 Execution of quantitative study

This section explains the quantitative aspect of the work which comprises the survey and archival research strategies. The survey strategy was executed with questionnaire data collection while the archival research was executed with company documentation in which case the financial ratios of firms were downloaded from a financial database

6.7.1 Questionnaire data

6.7.1.1 Pilot study for questionnaire

The criteria used to select respondents is as explained in subsection 6.6.1. The themes that resulted from analysing the qualitative data (see section 7.2) were used to develop a preliminary questionnaire to determine how relevant each identified variable/criterion is in determining (in)solvency of construction firms. Where multi-scale beyond two points was applicable, a Likert scale of one to five points was used. This preliminary questionnaire was used as a pilot study with the aim of evaluating its relevance/correctness, complexity, length and layout before being sent out to a wider set of target respondents. An initial pilot study was conducted using 20 colleagues with experience in the construction industry. Then a final pilot study was conducted with 11 volunteer respondents from the interviewees (two and three from existing big and MSM construction firms respectively, and two and four from failed/insolvent big and MSM construction firms respectively).

The key feedback was to reduce the number of questions which initially stood at 200. This was cut back to 111. There were also suggestions to rephrase some questions in order to make them more concise. Another vital feedback was the notification that some questions were not valid and should be removed. An example of identified invalid question was 'what percentage of the directors are/were married?' All feedback/suggestions were diligently implemented.

6.7.1.2 Sampling and execution of survey (questionnaire) strategy

To conduct the actual survey for the research, the sampling strategies used for the qualitative study (see subsection 6.6.1) were repeated but extended to reach more potential respondents. For existing firms, an onerous exercise of identifying 1200 firm directors/owners was done using random sampling and a hard copy questionnaire, addressed to each target respondent and attached with an official return envelope, was posted to each of them. A proportion of 20% for large firms and 80% for MSM firms was used in recognition of the skewed nature of the construction industry in terms of firm size. Where the target respondent email was traceable, a link to an online version of the questionnaire was sent to him/her via email instead

For failed firms, having realised that the random sampling using FAME search was relatively less effective, an extra onerous exercise of searching for present contacts of 1428 directors/owners of failed construction firms was done. All failed large construction firms whose directors could be traced were selected, resulting in a representation of 128 firms of the 146 returned in the search. The remaining 1300 contacts gotten were for failed MSM construction firms. The convenience sampling for the survey was extended by involving all construction apprentices in the college. Further, using the author and his college colleagues' links with other lecturers in other colleges, the questionnaires were also distributed to construction apprentices of another three colleges to pass on to potential respondents. In addition, all author's contacts, old or new, who have worked in the construction industry were contacted for help to give the questionnaire to any fitting respondents. The convenience sampling resulted in the distribution of another over 350 questionnaires, totalling around 3028 questionnaires distributed for failed and existing firms. A hard copy questionnaire addressed to each target respondent and attached with an official return envelope was sent to all identified potential respondents. Where the potential respondent was not known, as was with many convenience sample respondents, the questionnaire was addressed to ‘respondent’.

Overall, a total of 553 questionnaires were returned representing approximately an 18.3% return rate. This consisted of 284 and 269 from respondents of existing and failed construction firms respectively. Only 7.7% and 11.9% of the of the questionnaires from respondents of existing and failed construction firms respectively were for large construction firms. A preliminary assessment of the questionnaires revealed that in some very few cases, more than one (usually two) questionnaires were completed by respondents of a particular firm. In such cases, the average values of the values chosen by the two respondents were used to create a new questionnaire response for such firms. After doing this, the total usable number of questionnaires were 272 and 259 for existing and failed construction firms respectively.

The variables used in the questionnaire along with the theory bounding them are presented in Table 6.6. Only the variables in sections C to G of the questionnaire are presented in the table because sections A and B are about the demographics of the respondent and firm; they were not involved in the analysis. A complete sample of the distributed questionnaire for failed and existing firms is given in Appendix A.

Table 6.6: Questionnaire variables created from analysing the qualitative study and the theories bounding them

Section	Variables developed from analysing qualitative study	Theory
	Section C	
C	**Senior management and finance questions**	
C1.	The firm is/was owned by a single person	Upper echelon theory
C2.	The owner is/was the same person as the chief executive (CEO)/president/ Managing Director (MD) of the firm	Upper echelon theory
C3.	The firm has/had a board of directors	Upper echelon theory
C4.	If yes, how many directors does/did the firm have?	Upper echelon theory
C5.	The firm took over another firm at some point in time	Mintzberg’s 5Ps Perspective

Section	Variables developed from analysing qualitative study	Theory
C6.	If yes, was the takeover as a result of financial or other types of distress?	Mintzberg's 5Ps perspective
C7.	The firm has/had a clear bidding strategy	Porter's perspective
C8.	There is/was a clear sub-contractor selection process	Mintzberg's 5Ps perspective
C9.	The firm has/had a long term strategic goal	Organization ecology
C10.	The firm is/was specialised in a particular trade or service	Organization ecology
C11.	Has the range of trade/services broadened over time	Adaptationist perspective/ Dynamic capabilities
C12.	The firm change its main specialisation of construction work (e.g. from public to private project, or from building residential homes to commercial stores, among others) at some point in time	Adaptationist perspective/ Dynamic capabilities
C13.	The owner is/was on a fixed salary	Upper echelon theory
C14.	There is/was a dedicated financial director	Upper echelon theory
C15.	The financial director is/was performing another role at the same time	Upper echelon theory
C16.	The company account is/was clearly separated from any personal accounts	Upper echelon theory
C17.	Was account management fully computerised	Adaptationist perspective/ Dynamic capabilities
C18.	The firm consistently run/ran negative cash flow	Mintzberg's 5Ps perspective
C19.	The firm went through an expansion programme less than two years ago or within two years before closing down	Dynamic capabilities
	Section D	
D	**Proportion of firms' professionals with high qualifications/skills and involvement**	
D1.	Percentage of passive members on the board of directors	Upper echelon theory
D2.	Percentage of directors that worked in the firm	Upper echelon theory
D3.	Percentage of directors that had construction background	Upper echelon theory
D4.	Percentage of directors that had management/administrative background	Upper echelon theory
D5.	Percentage of directors educated to at least a degree level	Upper echelon theory
D6.	Percentage of personnel educated to at least a degree level	Upper echelon theory
D7.	Percentage of works usually subcontracted during projects	Mintzberg's 5Ps perspective / Adaptationist perspective
D8.	Percentage of successful bids	Adaptationist perspective

Section	Variables developed from analysing qualitative study	Theory
D9.	Percentage of firm's earnings invested in properties	Mintzberg's 5Ps perspective / Adaptationist perspective
D10.	Percentage of firm's earnings used in construction operations	Mintzberg's 5Ps perspective / Adaptationist perspective
D11.	Percentage of professional workers that were registered with professional bodies	Strategy theory
	Section E	
E	**The effect of external, industrial and firm characteristic factors**	
E1.	The 2008 global financial crises [Economic recession(s)]	Organization ecology/ Porter's perspective
E2.	High immigration levels in the UK	Organisation ecology/ Porter's perspective
E3.	Influx of firms into the industry, (from across the country and outside the country)	Porter's perspective
E4.	Fluctuation in construction material costs	Porter's perspective
E5.	Construction industry culture	Porter's perspective/ Organization ecology
E6.	Construction industry environmental sustainability agenda	Adaptationist perspective/ Dynamic capabilities
E7.	Type/Quality of workforce available for employment	Organization ecology
E8.	Newness [i.e. how did newness (first four years) affect the performance of the firm in its early years?]	Adaptationist perspective
E9.	The company size	Adaptationist perspective
E10.	Fraud (if fraud ever happened, how it affected the firm?)	-
E11.	Natural disasters (whether directly on the firm or its projects)	Organization ecology
	Section F	
F	**Frequency of occurrence of some project related factors**	
F1.	Very late collection of payment for completed works	Organizational theory
F2.	Unsuccessful collection of payment for completed works	Resource based view
F3.	Get cash-strapped on projects (cash flow)	Resource based view
F4.	Reach debt limit with bank/financier	Resource based view
F5.	Renegotiate loan terms	Resource based view
F6.	Make profit on projects	Mintzberg's 5Ps perspective

Section	Variables developed from analysing qualitative study	Theory
F7.	Produce complete financial statements	Mintzberg's 5Ps perspective
F8.	Bid for jobs outside firm's speciality	Adaptationist perspective
F9.	Executed project cost more than the bidding price used to win contract	Resource based view
F10.	Submit very low bids because of fierce competition	Adaptationist perspective
F11.	Rely on government projects	Mintzberg's 5Ps perspective
F12.	Rely on private projects	Mintzberg's 5Ps perspective
F13.	Firm win major bids it submitted	Porter's perspective
F14.	Firm completes project within stipulated time frame	Resource based view
F15.	Firm completes project within bidding budget	Resource based view
F16.	Firm executes project to time and cost without conflict	Resource based view
F17.	Internal conflict arises within the firm	Adaptationist perspective
F18.	Internal conflict within the organisation gets uncomplicatedly resolved	Adaptationist perspective
F19.	Firm gets project through referral from another customer	-
F20.	Expansion of firm	Dynamic capabilities
F21.	Conflicts with clients on projects	Adaptationist perspective
F22.	Conflicts with subcontractor in terms of subcontractors not showing up, performing low-quality works.	Adaptationist perspective
F23.	Delay of payments to subcontractors.	Mintzberg's 5Ps perspective
F24.	Conflicts with other major parties on projects	Adaptationist perspective
F25.	Conflict /litigation/legal issues / dispute arise from completed projects	Adaptationist perspective
F26.	Losing out in conflict /litigation/legal issues /dispute cases	Adaptationist perspective
F27.	Customers offer repeat business	Porter's perspective
F28.	Repeated use of particular sub-contractor(s)	Porter's perspective
F29.	Materials are supplied to firm on credit	Porter's perspective/ Resource based view
F30.	Debts payment to suppliers are delayed	Resource based view
F31.	Legal advice sorted for contracts taken	Mintzberg's 5Ps perspective
F32.	Problems with labour cost	Resource based view
F33.	Execution of multiple projects simultaneously	Adaptationist perspective/ Dynamic capabilities/ Mintzberg's 5Ps perspective
F34.	Bid for projects outside main geographical area of comfort (city, county, region, among others)	Adaptationist perspective
F35.	Register accidents on its site	Mintzberg's 5Ps perspective

Section	Variables developed from analysing qualitative study	Theory
F36.	Replace key personnel	Dynamic capabilities/ Resource based view
F37.	Execute a highly financially challenging project	Resource based view
	Section G	
G	**The characteristics and performance level of the firm, its management and its staff**	
G1.	Enthusiasm of the project management team	Upper echelon theory
G2.	Level of overall competence of top management team	Upper echelon theory
G3.	The willingness of the top management team to take risk	Upper echelon theory
G4.	The motivation of the CEO/directors	Upper echelon theory
G5.	The tolerance of the CEO	Upper echelon theory
G6.	The decisiveness of the CEO/directors	Upper echelon theory
G7.	Leadership support of CEO/directors to employees	Upper echelon theory
G8.	The creativity/innovation of the CEO/directors	Upper echelon theory
G9.	The integrity/transparency of the CEO/directors	Upper echelon theory
G10.	The flexibility of the CEO/directors	Upper echelon theory
G11.	The reliability/dependability of the CEO/directors	Upper echelon theory
G12.	The construction industry knowledge of the CEO/directors of the firm	Upper echelon theory/ Adaptationist perspective
G13.	The CEO's/directors' 'response to feedback'	Upper echelon theory
G14.	Commitment of project management team	Upper echelon theory
G15.	Level of firm's response to market change	Porter's perspective
G16.	The effectiveness of the financial director	Upper echelon theory
G17.	The profit levels of the firm	Resource based view
G18.	The liquidity level of the firm	Resource based view
G19.	Firm's reception to latest technologies	Dynamic capabilities

6.7.2 Company documentation data

Using the FAME financial database, the financial data of the 272 and 259 existing and failed construction firms with usable questionnaire data were downloaded. This means the financial data of a total of 531 construction firms were downloaded. Only the data of the final year of failed construction firms and the most recent financial data of existing firms were downloaded for use in building the construction firms insolvency prediction model

(CF-IPM) of the research. A typical financial ratio section of the financial data of all categories of sample construction firms are given in Table 6.7

Table 6.7: Typical financial ratios section of the financial statement of the sample construction firms

Financial ratios category	**Financial ratios (variable) name**	**Big existing firms**	**Big failed firms**	**MSM Existing firms**	**MSM failed firms**
Profitability ratios	Return on Shareholders Funds (%)	22.56	6.07	2.38	-9.52
	Return on Capital Employed (%)	17.95	6.07	2.38	-9.52
	Return on Total Assets (%)	14.88	2.12	2.37	-4.39
	Profit margin (%)	18.84	0.55	87.72	n.s.
	Gross margin (%)	21.67	10.15		n.s.
	Berry ratio	2.70	1.06		
	EBIT margin (%)	19.95	0.57	86.21	n.s.
	EBITDA margin (%)	21.26	0.57		n.s.
Operational ratios	Net Assets Turnover	0.95	11.09	0.03	n.s.
	Fixed Assets Turnover	2.34	n.s.	0.03	n.s.
	Interest Cover	15.27	21.75		
	Stock Turnover	2.67			
	Debtors Turnover	4.38	4.24	0.69	n.s.
	Debtor Collection (days)	83.43	86.07	531.91	n.s.
	Creditors Payment (days)	34.66	41.63	35.72	n.s.
Structure ratios	Current ratio	3.87	1.54	38.19	0.57
	Liquidity ratio	2.14	1.54	38.19	0.57
	Shareholders liquidity ratio	3.90			
	Solvency ratio (Asset based) (%)	65.98	34.99	99.74	46.11
	Solvency ratio (Liability based) (%)	n.s.	53.82	n.s.	85.57
	Asset Cover	5.91			
	Gearing (%)	27.41	32.89		

Financial ratios category	Financial ratios (variable) name	Big existing firms	Big failed firms	MSM Existing firms	MSM failed firms
Per employee ratios	Profit per employee (unit)	65,255	643	493	-444
	Turnover per employee (unit)	346,414	117,589	562	n.s.
	Salaries/Turnover	16.07	25.41	11.30	n.s.
	Average Remuneration per employee (unit)	55,655	29,884	64	7
	Shareholders' Funds per employee (unit)	n.s.	10,605	20,746	4,666
	Working Capital per employee (unit)	176,179	14,318	764	-5,330
	Total Assets per employee (unit)	438,448	30,310	20,801	10,119

EBIT: Earnings before interest and tax

n.s.: not available/applicable.

6.8 Validity and reliability

A major step towards validity was the pilot studies for the interview and the questionnaire as explained in subsections 6.6.2 and 6.7.1.1 respectively. For the interviews, the respondents unanimously agreed that the initial question asked in the interview was valid. Each respondent also agreed with the validity of the follow up questions that were asked. For the questionnaire, a few invalid questions were identified by the respondents while all questions in the final questionnaire were unanimously agreed to be valid. Question validity is vital to validity of research according to Bailey (1994).

The second validation step taken with the interview data was to ensure coding validity. This was done by having another experienced researcher, my second supervisor in this case, code the interview data (i.e. carry out a thematic analysis) independently. His codes/themes were subsequently compared with mine and there was reasonable agreement between the results. This code/theme validation process remains one of the most common and acceptable check of validity of interview data (Bailey 1994; Mays and Pope 1995; Rolfe 2006).

The predictive validity test, which is a criterion related validity test, was used as the second validation method for the questionnaire data and the first and only validation method for the archival data (i.e. the final ratios). This test is commonly used for measurements (Van Dyne and LePine 1998) e.g. questionnaire ratings and financial ratios in this case. It is used to predict future event or outcome of interest. The process involved splitting the data and putting the questionnaire and archival data of some of the sample firms together to develop CF-IPMs. The data of the remaining sample firms was then fed into the CF-IPMs to predict if the firms were failing or existing. The CF-IPMs had great prediction accuracy (see table 8.21 in subsection 8.55), thus confirming the validity of the data. This is the main aim of the research work as given section 1.7. More on data split can be found in section 8.3 while more on presentation of the prediction results can be found in subsection 8.5.3

Reliability of research refers to the degree to which the research is consistent or repeatable. In terms of data collection, the fewer the number of respondents, the more tendency the data collected will vary from that of another researcher because each researcher might have encountered entirely separate respondents. In essence, the higher the number of respondents, the higher the reliability of data.

Marshall, Cardon, Poddar, and Fontenot (2013) in their comprehensive review of sample sizes in qualitative research, were able to establish that saturation is reached in interview data long before interviewing 30 respondents. Guest *et al.* (2006) had earlier been able to identify the required number of respondents for saturation in qualitative research to be 12 respondents. Saturation is the point at which no new information comes out of subsequent interviews and does represent a point at which reliability can be said to have been met. The conduction of interviews for 13 and 14 respondents (both greater than 12) to investigate the construction firms survival and failure phenomenon respectively (see tables 6.3 and 6.5 respectively) thus make the interview data in the research reliable.

One of the ways of measuring reliability for questionnaire surveys is by checking the relative response rate (Fincham, 2008). Baruch and Holtom's (2008) comprehensive review of 490 organization research studies that used questionnaire revealed the standard response rate for studies that focused on organizations (e.g. construction firms as in this case) was 35.7% with a standard deviation of 18.8%. This indicates a minimum response rate of 17.4%. The response rate of 18.3% achieved in the research thus gives a comparable data reliability with other organization research studies. The other measure used to check the reliability of the questionnaire data was to run a reliability test using the Cronbach’s alpha analysis. A detailed explanation of this can be found in subsection 7.3.1

6.9 Chapter summary

The methodological positions of CF-IPM studies were reviewed in this chapter with more focus on paradigm, epistemology and ontology than other areas. The review exposed the blind followership act of CF-IPM studies where they restrict themselves to the use of positivism paradigm, realism ontology and objectivism epistemology. This singular dimension approach, which normally involves the exclusive of financial ratios, does not fully represent the insolvency situation of construction firms as highlighted in various studies; and due to the dynamism of the construction industry. The need to involve social factors, which can mainly be considered through a subjective approach, and the need to talk to subjects to understand the timely dynamics of the construction industry, both call strongly for the adoption of the subjective epistemology in CF-IPM studies.

The research thus combines the subjective and objective epistemologies, using the pragmatism paradigm which allows a research to use or combine any set of research positions as long as they will best answer the research question. The subjective epistemology was established using the multiple case study approach. This was done using the storytelling strategy which was executed with unstructured interviews. The objective epistemology was established using the survey strategy executed with questionnaires.

The unit of analysis is construction firms. The target respondents were owners and directors of large and MSM construction firms. The random sampling method alone, based on search results from FAME database, was used to identify potential respondents for existing construction firms. The method was combined with the convenience and snowballing sampling methods for respondents of failed/insolvent firms. The variables used in the questionnaire was formulated based on literature review and result of the qualitative study (i.e. unstructured interviews), depicting a facilitation approach. A total of 272 and 259 (total = 531) usable questionnaires for existing and failed construction firms respectively were gotten. The financial data of the 531 firms were subsequently downloaded from FAME.

Chapter seven contains a presentation of the types of analyses carried out on the qualitative and quantitative data. These included narrative and thematic for qualitative data, as well as reliability and factor analysis among others for quantitative data.

CHAPTER SEVEN

7.0 DATA ANALYSIS AND RESULTING VARIABLES FOR MODEL DEVELOPMENT

7.1 Chapter introduction

Data collected through the unstructured interview, questionnaire and company documentation methods (see chapter six) are analysed in this chapter to create the qualitative and quantitative variables for the construction firms insolvency prediction model (CF-IPM) to be developed in the next chapter. Section 7.2 is used to explain how narrative and thematic analyses were used to analyse the qualitative data collected through unstructured interview using the storytelling method (see section 6.6). The resulting themes were used to create questionnaire variables which have already been presented in table 6.6 of subsection 6.7.1. Section 7.3 is used to present various analyses which eventually produced the initial variables for the CF-IPM development process. Subsection 7.3.1 is used to present the reliability analysis of the questionnaire variables and the results. Subsection 7.3.2 is used to present how the financial ratios that will be used as quantitative variables for the research's CF-IPM were chosen based on being applicable to micro, small and medium (MSM) as well as large construction firms, thereby satisfying one of the objectives of the research. Subsection 7.3.3 is used to present how the sample matching method and synthetic minority over-sampling technique (SMOTE) algorithm were used to oversample financial and questionnaire data respectively from 531 construction firms to 1062 construction firms. Subsection 7.3.4 is a description of dimension reduction of the questionnaire variables. Section 7.4 provides a summary of the results of the analyses through a table showing all the 24 quantitative and qualitative variables created for the CF-IPM development process. Section 7.5 is a summary of the chapter

7.2 Qualitative data analysis

There are a number of approaches to analysing qualitative data, each approach stemming from different traditions. The research used the narrative and thematic qualitative analyses to analyse the qualitative data collected through unstructured interview using the storytelling method (see section 6.6). The narrative analysis, which is the usually employed technique for storytelling was used first, but in a secondary manner (Saunders, Lewis and Thornhill, 2000). In analysing and interpreting each respondent's stories, the transcripts were carefully read and each one was disaggregated into a number of recognisable insolvency episodes (D'hondt, 1994).

To identify factors and variables affecting (in)solvency of construction firms, thematic analysis was subsequently performed on all the episodes (Saunders, Lewis and Thornhill, 2000) using the Nvivo software. Both prior categories and new categories were used and developed respectively during the thematic analysis. Prior categories refer to issues already identified from the literature review of construction firms' insolvency studies (see section 3.4 of chapter three) while any issue identified during reading through the episodes were also used to construct conceptual categories which characterised major themes. The Nvivo software word frequency search was also used to create themes. Examples of coding of prior and new themes and the respondents' statements they were taken from are presented in Table 7.1.

In developing the themes, the transcripts were read repeatedly and discussed with supervisors, who also read them separately, in relation to both prior and newly constructed categories. Extra effort was made to maintain awareness of the effect of research process on the stories obtained during the interpretation and analysis of obtained data. It is acknowledged that many components of the research process such as respondent's talkative ability, command of interview/story language (i.e. English language), level of experience, social class, among others, may have had an effect on the eventual output. The findings are thus taken to be a construction process between the researchers and the respondents, as not representing a single truth, but instead as some possible stories of many potential stories. The resulting themes and sub-themes, commonly identified by the researcher and the supervisory team, were subsequently used as variables in the questionnaire used in the research, as presented in Table 6.6 under subsection 6.7.1 in chapter six.

Table 7.1: Example of coding from prior and new themes and the respondents' statements they are taken from

Theme	Prior or new	Statements	Respondent type*
New entrant's threats (Porters theory)	New	'The works dried out because people now prefer to give the jobs to some European **immigrants** that will do a shoddy job for a token.'	Respondent 1 (failed micro firm)
Collection of receivables	Prior and new	Construction is very interesting. You bring your stuff and workers in, get the job done and get paid. Easy money… But I stopped because	Respondent 8 (failed micro firm)

Theme	Prior or new	Statements	Respondent type*
		people don't pay up. You make several fruitless efforts that even cost you money.	
Conflict management capability and Legal cost	New	But I stopped because people don't pay up… **And they take you to court** if you dismantle the job despite you will incur losses on that.	Respondent 5 (failed small firm)
Over-reliance on accounting books to make decision	Prior and new	We made took our time and always consulted our books before making decisions. In fact, we ensured almost **no financial decision was taken without checking our account books**	Respondent 13 (failed medium firm)
Sustainability issue	Prior and new	Many people don't know what they want. **They want you to use only environmental friendly stuff for them, yet they also want the cheapest price.** They want to get what they don't want to pay for	Respondent 3 (Existing medium firm)
Strategy as plan (Mintzberg's 5 Ps), Economic recession	Prior	I understand property investment and **always buy houses and lands and sell them later.** Brother, this brings more money to do the building [i.e. construction]. The stupid **problem with economy [recession] caused all my property to go down [i.e. devalue]**. Brother, why is America problem our problem (hisses).	Respondent 17 [Existing micro firm (this firm was recovering according to the owner]
		The management invested too much in properties and the company incurred many losses during the recession... Although I strived to convince them against some investments, they wouldn't listen	Respondent 2 (failed large firm)

* Note that name and firm of the respondent cannot be disclosed for confidentiality reasons

7.3 Quantitative data analysis

7.3.1 Reliability analysis

As recommended by many social scientists (Field, 2009; George and Mallery, 2003; Nunnally and Bernstein, 1994; Spector, 1992; among others), the research used the Cronbach's alpha coefficient test to examine the reliability of the questionnaire data. Mathematically, according to Oyedele, (2013), Cronbach's alpha is written as

$$\alpha = \frac{N^2 \overline{COV}}{\sum S^2_{criteria} + \sum COV_{criteria}}$$

The goal of the test was to check the consistency of the obtained data to establish if the variables and their associated Likert scale are really measuring the construct they were intended to measure (Field, 2009). The construct, in this case, is the title given to each section/group of variables as related to the failure/survival of construction firms. Cronbach's alpha coefficient value ranges from 0 to 1 and as a thumb rule, 0.7 is suggested as the lowest acceptable score and 0.8 as an indication of good internal consistency, 0.9 and above represent high consistency (George and Mallery 2003). Table 7.2 presents the Cronbach's alpha coefficient test results gotten from SPSS. The reliability test was run and the overall Cronbach's alpha coefficient gotten for sections D, E, F and G variables were more than 0.8 (see details in Table 7.2), depicting good internal consistency of the data. This makes them useful for the research. The group C variables, however, returned a Cronbach's alpha coefficient of 0.523, a value much lesser than the lowest acceptable score. For this reason, the group C variables were not included in the development of the construction firms insolvency prediction model (CF-IPM) of the research.

Table 7.2: The questionnaire construction firms' (in)solvency variables and associated reliability analysis scores

Section	Variables per section	Cronbach's alpha if item deleted
	Section C	
C	**Senior management and finance questions**	
	Overall Cronbach's alpha coefficient for section G variables = 0.523	
C1.	The firm is/was owned by a single person	0.518
C2.	The owner is/was the same person as the chief executive (CEO)/president/ Managing Director (MD) of the firm	0.245
C3.	The firm has/had a board of directors	0.559
C4.	If yes, how many directors does/did the firm have?	0.278
C5.	The firm took over another firm at some point in time	0.209
C6.	If yes, was the takeover as a result of financial or other types of distress?	0.359

Section	**Variables per section**	**Cronbach's alpha if item deleted**
C7.	The firm has/had a clear bidding strategy	0.294
C8.	There is/was a clear sub-contractor selection process	0.440
C9.	The firm has/had a long term strategic goal	0.247
C10.	The firm is/was specialised in a particular trade or service	0.384
C11.	Has the range of trade/services broadened over time	0.594
C12.	The firm change its main specialisation of construction work (e.g. from public to private project, or from building residential homes to commercial stores, among others) at some point in time	0.231
C13.	The owner is/was on a fixed salary	0.482
C14.	There is/was a dedicated financial director	0.328
C15.	The financial director is/was performing another role at the same time	0.224
C16.	The company account is/was clearly separated from any personal accounts	0.364
C17.	Was account management fully computerised	0.311
C18.	The firm consistently run/ran negative cash flow	0.431
C19.	The firm went through an expansion programme less than two years ago or within two years before closing down	0.298
	Section D	
D	**Proportion of firms' professionals with high qualifications/skills and involvement**	
	Overall Cronbach's alpha coefficient for section G variables = 0.886	
D1.	Percentage of passive members on the board of directors	0.839
D2.	Percentage of directors that worked in the firm	0.823
D3.	Percentage of directors that had construction background	0.808
D4.	Percentage of directors that had management/administrative background	0.863
D5.	Percentage of directors educated to at least a degree level	0.852
D6.	Percentage of personnel educated to at least a degree level	0.846
D7.	Percentage of works usually subcontracted during projects	0.787
D8.	Percentage of successful bids	0.804
D9.	Percentage of firm's earnings invested in properties	0.860
D10.	Percentage of firm's earnings used in construction operations	0.836
D11.	Percentage of professional workers that were registered with professional bodies	0.870

Section	Variables per section	Cronbach's alpha if item deleted
	Section E	
E	**The effect of external, industrial and firm characteristic factors**	
	Overall Cronbach's alpha coefficient for section E variables = 0.892	
E1.	The 2008 global financial crises [Economic recession(s)]	0.863
E2.	High immigration levels in the UK	0.879
E3.	Influx of firms into the industry, (from across the country and outside the country)	0.871
E4.	Fluctuation in construction material costs	0.862
E5.	Construction industry culture	0.884
E6.	Construction industry environmental sustainability agenda	0.889
E7.	Type/Quality of workforce available for employment	0.819
E8.	Newness [i.e. how did newness (first four years) affect the performance of the firm in its early years?]	0.871
E9.	The company size	0.876
E10.	Fraud (if fraud ever happened, how it affected the firm?)	0.853
E11.	Natural disasters (whether directly on the firm or its projects)	0.877
	Section F	
F	**Frequency of occurrence of some project related factors**	
	Overall Cronbach's alpha coefficient for section F variables =0.822	
F1.	Very late collection of payment for completed works	0.786
F2.	Unsuccessful collection of payment for completed works	0.786
F3.	Get cash-strapped on projects (cash flow)	0.787
F4.	Reach debt limit with bank/financier	0.790
F5.	Renegotiate loan terms	0.805
F6.	Make profit on projects	0.796
F7.	Produce complete financial statements	0.795
F8.	Bid for jobs outside firm's speciality	0.803

Section	Variables per section	Cronbach's alpha if item deleted
F9.	Executed project cost more than the bidding price used to win contract	0.808
F10.	Submit very low bids because of fierce competition	0.798
F11.	Rely on government projects	0.803
F12.	Rely on private projects	0.806
F13.	Firm win major bids it submitted	0.810
F14.	Firm completes project within stipulated time frame	0.782
F15.	Firm completes project within bidding budget	0.789
F16.	Firm executes project to time and cost without conflict	0.790
F17.	Internal conflict arises within the firm	0.803
F18.	Internal conflict within the organisation gets uncomplicatedly resolved	0.772
F19.	Firm gets project through referral from another customer	0.772
F20.	Expansion of firm	0.807
F21.	Conflicts with clients on projects	0.797
F22.	Conflicts with subcontractor in terms of subcontractors not showing up, performing low-quality works.	0.791
F23.	Delay of payments to subcontractors.	0.809
F24.	Conflicts with other major parties on projects	0.805
F25.	Conflict /litigation/legal issues / dispute arise from completed projects	0.813
F26.	Losing out in conflict /litigation/legal issues /dispute cases	0.803
F27.	Customers offer repeat business	0.789
F28.	Repeated use of particular sub-contractor(s)	0.783
F29.	Materials are supplied to firm on credit	0.801
F30.	Debts payment to suppliers are delayed	0.810
F31.	Legal advice sorted for contracts taken	0.806
F32.	Problems with labour cost	0.809
F33.	Execution of multiple projects simultaneously	0.773
F34.	Bid for projects outside main geographical area of comfort (city, county, region, among others)	0.809
F35.	Register accidents on its site	0.804
F36.	Replace key personnel	0.803
F37.	Execute a highly financially challenging project	0.793

Section	Variables per section	Cronbach's alpha if item deleted
	Section G	
G	**The characteristics and performance level of the firm, its management and its staff**	
	Overall Cronbach's alpha coefficient for section G variables = 0.925	
G1.	Enthusiasm of the project management team	0.907
G2.	Level of overall competence of top management team	0.866
G3.	The willingness of the top management team to take risk	0.876
G4.	The motivation of the CEO/directors	0.868
G5.	The tolerance of the CEO	0.881
G6.	The decisiveness of the CEO/directors	0.872
G7.	Leadership support of CEO/directors to employees	0.873
G8.	The creativity/innovation of the CEO/directors	0.866
G9.	The integrity/transparency of the CEO/directors	0.872
G10.	The flexibility of the CEO/directors	0.870
G11.	The reliability/dependability of the CEO/directors	0.884
G12.	The construction industry knowledge of the CEO/directors of the firm	0.882
G13.	The CEO's/directors' 'response to feedback'	0.916
G14.	Commitment of project management team	0.879
G15.	Level of firm's response to market change	0.872
G16.	The effectiveness of the financial director	0.919
G17.	The profit levels of the firm	0.922
G18.	The liquidity level of the firm	0.878
G19.	Firm's reception to latest technologies	0.862

To check if all the variables are contributing to the internal consistency of the data, the 'Cronbach's alpha if item deleted', located in column three of Table 7.2 is further investigated. A variable that is not contributing to the overall reliability of a section's set of variables will normally have a higher associated 'Cronbach's alpha if item deleted' value than the data's overall Cronbach's alpha coefficient (Field, 2009). This higher value depicts that if the variable with the value is deleted, the overall reliability of the data will increase

(Field, 2009). In this context, all the variables in sections D, E, F and G contribute positively to the sections' reliability. On the other hand, variables C3 and C11 with 'Cronbach's alpha if item deleted' scores of 0.559 and 0.594 respectively, which are each greater than section C's overall Cronbach's alpha coefficient of 0.523, do not contribute to the reliability of the group. This, however, does not really matter in this case as the section has a poor overall Cronbach's alpha coefficient and its variables will not be involved in model building as clarified earlier

7.3.2 The financial ratios used as quantitative variables

The initial financial ratio selection was based on the second objective listed in section 1.7 (aims and objectives) of chapter one which is to identify the quantitative (financial) variables that are commonly reported by micro, SME and large construction firms. This was very important as the data clearly indicated a pattern whereby micro, small and medium (MSM) construction firms always omitted reporting certain financial ratios. The financial statements of the sample firms were carefully studied and the 11 ratios, out of the available 29 (see Table 6.7 in subsection 6.7.2 of chapter six), most reported by all categories of firms (i.e. large, MSM, failed and existing) were identified. These ratios are presented in Table 7.3 and are coded with the letter 'R' to represent ratio. Hence the quantitative variables are represented with letter 'R' in the research

Table 7.3: The ratios most reported by all categories of construction firms (i.e. large, MSM, failed and existing)

Financial ratios (variable) category	**Financial ratios (variable) name**	**Assigned (quantitative) variable code**
Profitability ratios	Return on Shareholders Funds (%)	R1
	Return on Capital Employed (%)	R2
	Return on Total Assets (%)	R3
Structure ratios	Current ratio	R4
	Liquidity ratio	R5
	Solvency ratio (Asset-based) (%)	R6

Per employee ratios	Profit per employee (unit)	R7
	Average Remuneration per employee (unit)	R8
	Shareholders Funds per employee (unit)	R9
	Working Capital per employee (unit)	R10
	Total Assets per employee (unit)	R11

7.3.3 Oversampling and pairing of financial and questionnaire variables

Financial data oversampling: Using the sample matching method, the financial data of another 531 construction firms, in addition to the previously downloaded 531 financial data, were downloaded to have a total sample size of 1062 construction firms comprising of 518 and 544 existing and failed construction firms respectively. This is a method of oversampling (see subsection 5.2.2 of chapter five). Sample matching is a process where, for each firm in a sample, a search is done for a firm with very similar characteristics and added to the sample. In this case, the characteristics used to match the sample construction firms include, trade specialism, turnover, number of employees, year of establishment, year of failure where applicable and number of directors. All of this information is available in the FAME database.

Questionnaire data oversampling: Using the ‘R’ software, the Synthetic Minority Over-Sampling Technique (SMOTE) algorithm was used to double (oversample) the number of questionnaire responses available from 531 to 1062, just as done with the financial data. The algorithm does not give a repeat of any of the existing data. Rather it studies the data pattern and creates new ones based on the study (Chawla, Bowyer and Hall, 2002). This means it creates data for the same number of large, MSM, failed and existing construction firms as present in the original data hence the oversampled data equally contained 518 and 544 existing and failed construction firms respectively as the original data.

Pairing financial and questionnaire data: The original questionnaire data of 531 construction firms and their corresponding financial data downloaded from FAME were paired to make each firm have questionnaire and financial variables (i.e. qualitative and quantitative variables respectively). The 531 SMOTE oversampled questionnaires were then carefully studied in comparison to the original 531 questionnaires to identify the ones that represented various groups, (i.e. MSM failed firms, MSM existing firms, large failed firms and large existing firms). Using the sample matching method for the financial data meant that the group each oversampled financial statement belonged to was readily known. The oversampled questionnaire data belonging to each group were subsequently randomly paired with oversampled financial data in the same group thereby *ensuring that financial and questionnaire data from different groups were never paired.* **Paired questionnaire and financial variables (i.e. qualitative and quantitative variables respectively) for a total of 1062 sample firms were thus available as data for developing CF-IPMs**

7.3.4 Factor analysis and the developed qualitative variables

An initial attempt to build models using the very many questionnaire variables combined with financial variables did not work out as the optimisation runs required for classification did not converge due to too many questionnaire variables. The decision was thus taken to reduce the number of questionnaire variables using ‘dimension reduction’ executed with factor analysis. The explorative factor analysis was carried out using the SPSS software. The ‘principal component’, ‘generalised least squares’, ‘maximum likelihood’ and ‘principal axis factoring’ methods were initially used in succession to extract the factors in an attempt to decide the right number of factors to be extracted. All methods resulted in a total number of 13 factors.

For the final factor extraction, the ‘maximum likelihood’ method and ‘direct oblimin’ oblique rotation were used as methods of factor extraction and rotation respectively. Having noticed that rotation did not converge with the default 25 iterations setting during the initial extractions, a value of 50 was entered for ‘maximum iterations for convergence’ in the rotation dialogue box. Kaiser-Meyer-Olkin (KMO) and Bartlett tests of sphericity measure of sampling adequacy were conducted to check the appropriateness of the data for factor analysis. Since the ‘Eigenvalue greater than one’ criterion has been established to be based on misemployment of the internal consistency reliability formula (Cliff, 1988), the scree plot and the initial application of four different extraction methods were used to decide that 13 factors were to be extracted. A value of 13 was thus entered in the number of ‘factors to extract:’ box for the analysis. To create new representative variables for each extracted factor, the ‘save as variables’ box was ticked in the scores dialogue box, with the default regression method selected to create the variables.

The result of the analysis produced values of 0.839 (above 0.5) and 0.00034965 (less than 0.05) were gotten for KMO and Bartlett tests of sphericity respectively, demonstrating that the data set is suitable for factor analysis and the sampling is adequate (Pallant, 2013). According to Pallant (2013), the closer the KMO value to one, the more the appropriate the use of factor analysis. Table 7.4 presents the details of the total variance of the 13 extracted factors.

Table 7.4: Total Variance Explained

Factor (i.e. new variables create)	**Initial Eigenvalues**			**Extraction Sums of Squared Loadings**			**Rotation Sums of Squared Loadings**[a]
	Total	**% of Variance**	**Cumulative %**	**Total**	**% of Variance**	**Cumulative %**	**Total**
1)	20.603	26.414	26.414	20.358	26.099	26.099	6.879
2)	11.308	14.498	40.911	10.947	14.034	40.134	8.300
3)	8.264	10.595	51.506	8.188	10.498	50.631	7.638
4)	6.373	8.171	59.677	6.174	7.916	58.547	10.983
5)	3.743	4.799	64.476	3.468	4.447	62.993	5.933
6)	3.587	4.599	69.075	3.423	4.389	67.382	4.050
7)	3.019	3.871	72.946	2.809	3.601	70.983	8.937
8)	2.318	2.971	75.917	1.974	2.531	73.515	10.433
9)	2.155	2.763	78.680	2.194	2.813	76.328	6.593
10)	1.540	1.974	80.654	1.374	1.761	78.090	7.057
11)	1.405	1.801	82.455	1.147	1.471	79.561	8.879
12)	1.273	1.632	84.088	1.120	1.435	80.996	5.736
13)	1.034	1.325	85.413	.915	1.173	82.169	7.238
Total					**82.168**		
Extraction Method: Maximum Likelihood.							
a. When factors are correlated, sums of squared loadings cannot be added to obtain a total variance.							

Although the 'Eigenvalue greater than one' criterion has been discredited (Cliff, 1988), the Eigenvalues (5th column) of all the extracted factors were greater than one except in the case of the 13th extracted factor, whose Eigenvalue was also very close to one. The extracted factors represented 82.168% of total variance (see the base of 6th column) as presented in Table 6.4; this portrays a good proportion of representation. As against the percentage of variance (6th column), the varimax rotated solution (8th column) produced values that portray a more evenly representation of the data by the extracted factors after redistribution, thereby giving more credence to the variance of the factors.

The pattern matrix table in the SPSS factor analysis result, which is used to select the variables representing each extracted factor, was used to produce Table 7.5. For factor grouping (i.e. to select variables representing an extracted factor), questionnaire variables with a factor loading of +0.3 and above or -0.3 and below were taken as part of the offspring of their principal factor (Child, 2006). No questionnaire variable in this analysis had a factor loading outside this range as evident in Table 7.5.

The questionnaire variables under each extracted factor are arranged in descending order in Table 7.5 based on the factor loading value, the variable with the highest factor loading value appearing in the first row of each extracted factor. This arrangement did not take the sign (positive or negative) of the factor loading value into consideration. However, only variables of the same sign can be, and were, taken as offspring/representing any one extracted factor. As such, offspring variables with the most common sign under a particular extracted factor were taken as variables contributing to that factor. For example, the offspring variables of 1st Extracted Factor are D5, D7, D2, F1 and D9 in that order (see Table 7.5). While D5, D7, D2, and D9 have negative factor loading values, F1 has a positive factor loading value and was thus considered **not** to be contributing to the 1st extracted factor. On the other hand, there are nine offspring of the second extracted factor, six with positive factor loading values and the remaining three with negative values; the three with negative factor loading values (a minority in this case) were thus not considered to be contributing to the extracted factor. All non-contributory offspring variables are given in *italics* font in Table 7.5. Further, any questionnaire variable that loaded significantly on more than one extracted factor was totally excluded i.e. not considered as part of any of the two or more extracted factors (Tabachnick and Fidell 2007). This is the case with D4 which loaded significantly on the third and fourth extracted factors. This questionnaire variable (i.e. D4) is thus given in italics font and underlined at the same time in Table 7.5.

Each extracted factor now represents a qualitative variable that will be used in the development process of the CF-IPMs. Each has been given a name based on the contributing constituent/offspring questionnaire variables (see column six of Table 7.5). Where all the contributing offspring questionnaire variables cannot be represented with a single name, a double-barreled name is used with the conjunction ‘and’. An example of this is the 4th Extracted Factor which is named ‘strategic issues and external relations’. In this case, some offspring represent strategic issues while others represent external relations. There are also a few instances where one of the offspring variables was not represented despite the use of a double barrelled name. Again, an example of this is the 4th extracted factor where the D1 variable (‘percentage of passive members on the board of directors’) does not fit the double-barreled extracted factor name: ‘strategic issues and external relations’. In such cases, nothing was done.

Lastly, there was a special case where the offspring of two of the extracted factors (third and twelfth) represent more or less the same thing hence they were given the same name which was separated with numbers: top management characteristics 1 and 2 for third and twelfth extracted factors respectively. The 13 extracted factors are finally coded Q1 to Q13, the letter ‘Q’ representing questionnaire (i.e. qualitative) variables.

Table 7.5: The extracted factors and the qualitative variables formulated from them

	Questionnaire variables used in the factor analysis	**Factor loadings**	**Percentage of variance**	**Eigenvalue**	**Extracted factor (qualitative variable) assigned name**	**Assigned variable code**
	Offspring variables of 1st Extracted Factor					
D5	Percentage of directors educated to at least a degree level	-0.724	26.099	20.358	Management knowledge and involvement	Q1
D7	Percentage of works usually subcontracted during projects	-0.677				
D2	Percentage of directors that worked in the firm	-0.515				
F1	*Very late collection of payment for completed works*	*0.448*				
D9	Percentage of firm's earnings invested in properties	-0.441				
	Offspring variables of 2nd Extracted Factor					
F15	Firm completes project within bidding budget	0.788	14.034	10.947	Construction Organization experience	Q2
F16	Firm executes project to time and cost without conflict	0.765				
F14	Firm completes project within stipulated time frame	0.681				
E8	Newness [i.e. how did newness (first four years) affect the performance of the firm in its early years?]	0.556				
F29	Materials are supplied to firm on credit	0.546				
E3	Influx of firms into the industry, (from across the country and outside the country)	0.505				
F11	*Rely on government projects*	*-0.374*				
F8	*Bid for jobs outside firm's speciality*	*-0.373*				

	Questionnaire variables used in the factor analysis	**Factor loadings**	**Percentage of variance**	**Eigenvalue**	**Extracted factor (qualitative variable) assigned name**	**Assigned variable code**
D6	*Percentage of personnel educated to at least a degree level*	*-0.368*				
	Offspring variables of 3rd Extracted Factor					
G10	The flexibility of the CEO/directors	0.892	10.498	8.188	Top management characteristics 1	Q3
G8	The creativity/innovation of the CEO/directors	0.874				
G7	Leadership support of CEO/directors to employees	0.843				
G2	Level of overall competence of top management team	0.693				
G1	Enthusiasm of the project management team	0.674				
G6	The decisiveness of the CEO/directors	0.559				
G4	The motivation of the CEO/directors	0.508				
F25	*Conflict /litigation/legal issues / dispute arise from completed projects*	*-0.429*				
D4	*Percentage of directors that had management/administrative background*	*0.348*				
	Offspring variables of 4th Extracted Factor					
F26	Losing out in conflict /litigation/legal issues /dispute cases	0.966				
F36	Replace key personnel	0.83				
F35	Register accidents on its site	0.653				
E2	High immigration levels in the UK	0.533				

	Questionnaire variables used in the factor analysis	**Factor loadings**	**Percentage of variance**	**Eigenvalue**	**Extracted factor (qualitative variable) assigned name**	**Assigned variable code**
E1	The 2008 global financial crises [Economic recession(s)]	0.519	7.916	6.174	Strategic and external issues	Q4
F23	Delay of payments to subcontractors.	0.467				
D4	*Percentage of directors that had management/administrative background*	*0.447*				
G9	*The integrity/transparency of the CEO/directors*	*-0.409*				
E10	*Fraud (if fraud ever happened, how it affected the firm?)*	*-0.393*				
E11	*Natural disasters (whether directly on the firm or its projects)*	*-0.39*				
D1	Percentage of passive members on the board of directors	0.373				
	Offspring variables of 5th Extracted Factor					
F27	Customers offer repeat business	0.913	4.447	3.468	Performance on projects	Q5
F33	Execution of multiple projects simultaneously	0.639				
F37	Execute a highly financially challenging project	0.446				
F19	Firm gets project through referral from another customer	0.35				
	Offspring variables of 6th Extracted Factor					
D8	Percentage of successful bids	0.873	4.389	3.423	Bidding issues	Q6
F13	Firm win major bids it submitted	0.45				
F34	Bid for projects outside main geographical area of comfort (city, county, region, among others)	0.438				

	Questionnaire variables used in the factor analysis	**Factor loadings**	**Percentage of variance**	**Eigenvalue**	**Extracted factor (qualitative variable) assigned name**	**Assigned variable code**
	Offspring variables of 7th Extracted Factor					
F22	Conflicts with subcontractor regarding subcontractors not showing up, performing low-quality works.	-0.864	3.601	2.809	Project related (external) conflict	Q7
F21	Conflicts with clients on projects	-0.849				
F24	Conflicts with other major parties on projects	-0.522				
	Offspring variables of 8th Extracted Factor					
D10	Percentage of firm's earnings used in construction operations	0.544	2.531	1.974	Finance and conflict related issues	Q8
F10	Submit very low bids because of fierce competition	0.524				
G19	*Firm's reception to latest technologies*	*-0.499*				
D11	*Percentage of professional workers that were registered with professional bodies*	*-0.495*				
F3	Get cash-strapped on projects (cash flow)	0.487				
F17	Internal conflict arises within the firm	0.47				
G18	The liquidity level of the firm	0.447				
F5	Renegotiate loan terms	0.42				
F4	Reach debt limit with bank/financier	0.408				
F12	Rely on private projects	0.328				

	Questionnaire variables used in the factor analysis	**Factor loadings**	**Percentage of variance**	**Eigenvalue**	**Extracted factor (qualitative variable) assigned name**	**Assigned variable code**
	Offspring variables of 9th Extracted Factor					
E6	Construction industry environmental sustainability agenda	0.762	2.813	2.194	External factors and management decisions	Q9
E5	Construction industry culture	0.719				
E9	The company size	0.661				
E4	Fluctuation in construction material costs	0.617				
E7	Type/Quality of workforce available for employment	0.499				
F20	Expansion of firm	0.425				
G3	The willingness of the top management team to take risk	0.411				
F18	Internal conflict within the organisation gets uncomplicatedly resolved	0.362				
	Offspring variables of 10th Extracted Factor					
F6	Make profit on projects	0.777	1.761	1.374	Profit issues	Q10
F9	Executed project cost more than the bidding price used to win contract	0.4				
	Offspring variables of 11th Extracted Factor					
G15	Level of firm's response to market change	-0.634				
D3	*Percentage of directors that had construction background*	*0.585*				
F32	Problems with labour cost	-0.472				
F30	Debts payment to suppliers are delayed	-0.468				

	Questionnaire variables used in the factor analysis	**Factor loadings**	**Percentage of variance**	**Eigenvalue**	**Extracted factor (qualitative variable) assigned name**	**Assigned variable code**
G17	The profit levels of the firm	-0.451	1.471	1.147	Cash flow and market issues	Q11
G16	The effectiveness of the financial director	-0.424				
	Offspring variables of 12th Extracted Factor					
G11	The reliability/dependability of the CEO/directors	0.829				
G13	The CEO's/directors' 'response to feedback'	0.645	1.435	1.120	Top management characteristics 2	Q12
G14	Commitment of project management team	0.599				
G5	The tolerance of the CEO	0.533				
F28	*Repeated use of particular sub-contractor(s)*	*-0.477*				
	Offspring variables of 13th Extracted Factor					
G12	The construction industry knowledge of the CEO/directors of the firm	0.811				
F2	Unsuccessful collection of payment for completed works	0.511	1.173	0.915	Industry contract/project knowledge	Q13
F31	Legal advice sorted for contracts taken	0.488				
F7	*Produce complete financial statements*	*-0.413*				

7.4 The complete set of quantitative and qualitative variables

The simple purpose of this section is to provide a summary result of all the analyses in this chapter. This is done with Table 7.6 which gives the complete set of quantitative (R1 to R11) and qualitative (Q1 to Q13) variables used in the CF-IPM development process in the next chapter

Table 7.6: Quantitative and qualitative variables used in the CF-IPM development process

Variable category	Serial number	Variable name	Assigned variable code
Quantitative variables	1.	Return on Shareholders' Funds (%)	R1
	2.	Return on Capital Employed (%)	R2
	3.	Return on Total Assets (%)	R3
	4.	Current ratio	R4
	5.	Liquidity ratio	R5
	6.	Solvency ratio (Asset-based) (%)	R6
	7.	Profit per employee (unit)	R7
	8.	Average Remuneration per employee (unit)	R8
	9.	Shareholders' Funds per employee (unit)	R9
	10.	Working Capital per employee (unit)	R10
	11.	Total Assets per employee (unit)	R11
Qualitative variables	12.	Management knowledge and involvement	Q1
	13.	Construction Organization experience	Q2
	14.	Top management characteristics 1	Q3
	15.	Strategic issues and external relations	Q4
	16.	Performance on projects	Q5
	17.	Bidding issues	Q6
	18.	Project related (external) conflict	Q7
	19.	Finance and conflict related issues	Q8
	20.	External factors and management decisions	Q9
	21.	Profit issues	Q10
	22.	Cash flow and market issues	Q11
	23.	Top management characteristics 2	Q12
	24.	Industry contract/project knowledge	Q13

7.5 Chapter summary

The data collected through unstructured interview (story telling method), questionnaire and company documentation methods were analysed in this chapter with the sole aim of creating the initial variables to be used in the development process of the proposed solution's CF-IPM. The stories were analysed using the commonly used narrative analysis with stories disaggregated into a number of recognisable insolvency episodes. To establish factors and variables affecting (in)solvency of construction firms, thematic analysis was subsequently performed on all the episodes using the Nvivo software. The resulting themes and sub-themes were subsequently used as variables in the questionnaire used in the proposed solution, as presented in Table 6.6 under subsection 6.7.1 in chapter six

A reliability analysis of the questionnaire variables was conducted. The result indicated that the scale used to measure section C variables of the questionnaire were not reliable hence they were excluded from any further analyses. Then the financial ratios from the financial statement of the sample firms, which were to be used as quantitative variables for the proposed solution's CF-IPM, were chosen based on being applicable to micro, small and medium (MSM) as well as large construction firms, thereby satisfying one of the objectives of the research. A total of 11 ratios were selected and coded with letter 'R' (i.e. R1 to R11). After the selection of the financial ratios (quantitative variables), the sample matching method and SMOTE algorithm were used to oversample financial and questionnaire data respectively from 531 construction firms' data to 1062 construction firms' data. Afterwards, the oversampled questionnaire data belonging to each group (i.e. MSM failed firms, MSM existing firms, large failed firms and large existing firms) were randomly assigned to oversampled financial data in the same group thereby ensuring that financial and questionnaire data from different groups were never paired. Paired questionnaire and financial variables (i.e. qualitative and quantitative variables respectively) for a total of 1062 sample firms were thus available as data for developing CF-IPMs

Finally, a dimension reduction analysis of the questionnaire variables was done using the factor analysis method. This was because the initial attempt to build CF-IPMs using the very many questionnaire variables combined with financial variables did not work out as the optimisations did not converge due to too many variables. The dimension reduction resulted in thirteen variables (Q1 to Q13) used as qualitative variables, with letter 'Q' representing questionnaire.

Chapter eight was used to present the process of developing the Big Data CF-IPMs of the research work. The process discussed include the set-up of Big Data Analytics platform, the variable selection process and the actual model development phase.

CHAPTER EIGHT

8.0 DEVELOPMENT OF BIG DATA PREDICTIVE ANALYTICS MODELS FOR CONSTRUCTION FIRMS

8.1 Chapter introduction

This chapter is a presentation of the details of how Big Data Analytics with some powerful predictive tools/algorithms were used to develop construction firms insolvency prediction models (CF-IPM) by first analysing and then employing some of the final quantitative and qualitative variables presented in chapter seven. Section 8.2 is an explanation of how the Big Data Analytics platform was set up with Apache Spark as the computation engine. The details include the cloud computing set up with the Amazon Elastic Compute Cloud and its 'spot instances' bought to hold data and execute computations. The details of the proportion of data used for training and testing the model are given in section 8.3. Section 8.4 is about how the variables used to develop the Big Data CF-IPMs were selected. Subsection 8.4.1 is a description of the 11 variable selection methods used, 8.4.2 is a detail of the implementation of the methods while 8.4.3 is a presentation of the result. The full details of how the Big Data CF-IPMS were developed are given in section 8.5. Subsection 8.5.1 is a description of the data pre-processing steps taken to improve the validity of the models. Some of the predictive tools/algorithms used to develop the Big Data CF-IPMs are lightly described in subsection 8.5.2. In subsection 8.5.3, the features of the Big Data CF-IPMs to be developed were explained to aid readers' understanding of the models. The 13 Big Data CF-IPMs developed were presented in subsection 8.5.4, each given a sub-subsection number (i.e. 8.5.4.1 to 8.5.4.13). The Big Data CF-IPMs include

1) Big Data Linear (Multiple) Discriminant Analysis CF-IPM
2) Big Data Quadratic (Multiple) Discriminant Analysis CF-IPM
3) Big Data Logistic Regression CF-IPM
4) Big Data Naïve Bayes CF-IPM
5) Big Data Support Vector Machine CF-IPM
6) Big Data K-Nearest Neighbour CF-IPM
7) Big Data Artificial Neural Network CF-IPM
8) Big Data Decision Tree CF-IPM
9) Big Data Random Forest CF-IPM
10) Big Data Bart Machine CF-IPM
11) Big Data Adaptive Boosting CF-IPM

12) Big Data Propositional Rule Learner CF-IPM

13) Big Data Kohonen CF-IPM

Subsection 8.5.5 is a presentation of the summary of the results of the Big Data CF-IPMs developed. Section 8.6 is a summary of the chapter.

8.2 Setting up the big data platform and the apache spark computation engine

The software chosen to develop the models is the ‘R’ programming language because it is very powerful and is less restricting in terms of operations compared to generic tools like SPSS and less generic tools like WEKA. Like any other programming language, ‘R’ requires a user to have knowledge of its codes for smooth operation.

Apache Spark is the framework or computation engine selected for the Big Data Analytics part of the proposed solution (see subsections 4.3.2 and 4.3.3 of chapter four). One of the reasons for its selection is its flexibility. Unlike Hadoop MapReduce that can only be used with Java, Apache Spark has an application programming interface (API) in the ‘R’ programming language, as well as in Java, Python, and Scala, which makes it very easy to run.

The data server used in the proposed solution is the Amazon Web Services Elastic Compute Cloud (AWS EC2) because of its provision for rationed usage that helps to reduce cost (in money) significantly. Rather than make full payment for the use of the server, the very cheap Spot Instances on the AWS EC2 were requested and used. Instances are virtual servers that have the capability to run applications. For the Amazon Machine Image of the instances, the ‘Ubuntu Server 14.04 LTS (HVM), SSD Volume Type’ option was selected. The ‘Instance’ option selected was ‘m4.large’ because of it was the cheapest option to include optimized-Elastic Block Store (EBS). This option allows volume size to be saved every time it is not in use. The saved portion then helps to provide a burst for 3000 seconds every time more volume than subscribed for is required. For ‘Instance Details’ configuration, six number of ‘Instances’ were requested with the maximum bid price set at £0.03 to reduce the possibility of losing the instance to a higher bidder.

Spark can execute iterations because it keeps data in its internal memory, using a storage model called resilient distributed datasets (RDD), for reuse. This means a large disk space is normally required to operate Spark . Hence the ‘Instance’ storage was set to 200gigabyte per ‘Instance’, and the ‘Magnetic’ option was chosen to reduce cost, though it reduces performance. The ‘Spot Instance’ was named ‘CF-IPM.BigDataCluster’. The ‘Configure Security Group’ option was set to OPEN_TO_ALL to make things easy because this option allows all IP addresses to access the instances. Restricting access could create troubles when working from different sources. When prompted for General Purpose volume, the ‘continue with Magnetic as boot volume’ option was selected. This action completed the setting up of the server/cloud for the Apache Spark to function as a Big Data Computation engine.

For installations, first, the ‘R’ programming language software was installed on the computer to be used. Then the Hadoop framework was installed because Apache Spark is partly dependent on it. With each of the six EC2 ‘Spot Instances’ taken as a node, five of the ‘Instances’ were run as Hadoop DataNodes and one as Hadoop NameNode. The Hadoop NameNode was subsequently configured as Spark Master. Finally, the data of the 1062 construction firms were uploaded to the five DataNodes making the system set for the Big Data Analytics development of CF-IPMs.

8.3 Details of model training and testing data

As proposed in Figure 6.3 in chapter six, the data for the proposed solution was divided into 70% and 30% for model (i.e. CF-IPM) training and testing respectively. The data in each case maintained the same ratio of existing to failed firms. The actual data contains 1062 construction firms including 544 and 518 existing and failed construction firms respectively representing a ratio of 51:49; this gives the required (very nearly) equal data dispersion. The details of the data composition for training and testing the models are given in Table 8.1. The ‘R’ codes used to execute the data split is given here.

```
split = sample.split(FacLogData$Status, SplitRatio = 0.7)
split
Train = subset(FacLogData, split == TRUE)
Test = subset(FacLogData, split == FALSE)
```

Table 8.1: Details of model training and test data for the CF-IPMs

	Existing construction firms	Failed construction firms	Total number of construction firms
Training data	381	363	744
Testing Data	163	155	318
Complete data	544	518	1062

8.4 Selection of variables

Variable selection is an important aspect of CF-IPM development. The most common variable selection method/technique in CF-IPM studies is the stepwise method (**see Table 5.1 in chapter five**) which normally employs the F-test or t-test. The stepwise technique has however been condemned for various reasons (Harrell, 2001) and is not a very sophisticated method. Other techniques used include regression, factor analysis, grey system theory, correlation analysis, among others. All these techniques give different but comparable results, and none is widely accepted as the best.

8.4.1 The techniques used

The proposed solution used some eleven sophisticated selection techniques in the 'R' programming language. Variable selection is considered as part of the model training process, as against model testing, hence only the training data was used in this process. A brief description of these techniques are as follows:

1. **Information gain:** This technique uses the decision tree to compare the information each independent variable contributes to making the correct classification (failed or existing construction firm class in this case) (Manning, Raghavan and Schutze, 2008).

2. **Kruskal test:** This technique is not computationally expensive (Ali Khan *et al.*, 2014). It measures the ability of each independent variable in separating two classes (failed and existing construction firm in this case). It yields a P-value for each variable. The closer the P value of an independent variable to zero, the more its predictive ability and importance.

3. **Minimum redundancy, maximum relevance filter (mRMR):** This is a very powerful technique which picks independent variables using two criteria simultaneously. The first is correlation where the independent variable that correlates most to the dependent variable is taken as the most importance. The second is multicollinearity; when two or more ‘collinear’ independent variables correlate to the dependent variable, it gives the less correlated variable(s) a negative sign and pushes it/them to the bottom of the pile (Acid, de Campos and Fernandez, 2011). This criterion guarantees the very important exclusion of multicollinearity in the selected variables

4. **Chi-squared:** This technique tests how uninfluential an independent variable is on the dependent variable by calculating an X^2 score. The higher the score, the less independent and the more important the independent variable is. Chi-squared is not a very accurate technique according to (Manning, Raghavan and Schutze, 2008).

5. **Gain ratio:** This is an extension of the information gain technique. It uses the C4.5 algorithm of decision tree to check how much an independent variable can separate the two classes (Gowda Karegowda, Manjunath and Jayaram, 2010)

6. **Analysis of variance (ANOVA) test:** ANOVA uses the sum of squared errors values to assess the extent of variation, between the two classes, each independent variable can explicate. The more variation a variable can explicate, the better it is.

7. **Cforest importance:** Cforest uses multiple (ensemble) decision trees and places each independent variable on each node before resampling to permute. The variables that show the best dependencies based on this process have the highest importance

8. **oneR:** This technique uses the association rule to associate each independent variable with the dependent variable and check which independent variable singularly makes the highest number of correct decisions or errors. oneR is more or less a univariate model developing process. The independent variable with the highest number of error classification is the least important.

9. **Relief:** This is a filter based variable selection technique which randomly selects the variables. It uses a small number of training and can only deal with a maximum of two classes (Rosario and Thangadurai 2015) as in this case.

10. **RF importance:** This technique uses the random forest algorithm to build multiple decision trees in order to rate the independent variables in order of importance.

11. **Symmetrical Uncertainty (SU):** This is a filter based approach that ranks the independent variables based on a calculated score. Like mRMR (Singh, Kushwaha and Vyas, 2014), its process include the consideration of multicollinearity, making it a strong selector (Senthamarai Kannan and Ramaraj 2010).

8.4.2 Implementation of the variable selection techniques

To implement the variable selection techniques discussed, the data was pre-processed as will be explained in section 8.5.1 and the necessary packages were installed before their libraries were unloaded. Libraries are computer packages required to run various algorithms (techniques) on the ‘R’ programing language. The following codes were used to install and unload the packages and libraries respectively, run the variable selection methods, plot the charts and export the tables

```
>install.packages("mlr")
>install.packages("mRMRe")
>install.packages("FSelector")
>install.packages("randomForestSRC")
>install.packages("kohonen")
>install.packages("party")
>library(mlr)
>library(mRMRe)
>library(FSelector)
>library(randomForestSRC)
>library(kohonen)
>library(party)
>fv=generateFilterValuesData(task, method = c("information.
gain", "kruskal.test", "mrmr", "chi.squared", "gain.ratio",
"anova.test", "cforest.importance", "oneR", "relief",
"rf.importance", "symmetrical.uncertainty"))
>fv$data
>plotFilterValuesGGVIS(fv)
```

```
>newobject=xtable(fv)
>print.xtable(newobject,type="html",file="VarSeltnFile.html"
)
```

8.4.3 Result of independent variables selection and the dependent variable

The raw variable scores from the analyses are given in Table 8.2 and Figures 8.1 to 8.11. The results in Table 8.2 are sectioned into quantitative and qualitative variables. Each set of variables are arranged sequentially (i.e. R1-R11 for quantitative variables and Q1-Q13 for qualitative variables). The charts in Figures 8.1 to 8.11 on the hand display bar plots of the independent variables arranged according to their scores for each the 11 selection techniques. The bars in the chart are an indication of the importance level attached to each variable by the associated selection method. It can be seen from Figure 8.3 that the mRMR selection method even assigned negative importance to some variables, showing that they may be of multicollinearity with some other variables. Since there is serious disagreement over which technique is best, two voting systems (Tables 8.3 and 8.4) were used to rank the variables while a combined ranking from the two systems (Table 8.5) was used for the final selection of the variables. Table 8.6 is a full presentation of the quantitative and qualitative variables selected to develop the CF-IPM. The voting and ranking systems are explained.

Percentage score ranking system: The score ranking system was implemented by adding the score assigned to each independent variable by all the selection techniques to give a total score. Since the scales of the scores used by the selection techniques were very different (see Table 8.2), the scores for the quantitative and qualitative variables could not be added directly to avoid bias. To ensure equal contribution from all techniques, the scores were converted to percentage under each selection technique. In essence, the total quantitative and qualitative variables scores were each taken as 100%. The percentage scores were then added to give a total score for each variable, and the rankings of the variables were done based on these total percentage scores (see Table 8.3). One advantage of this system is that it takes into consideration the difference in score between variables, under each technique, even if they are ranked next to each other. An example of such case is in the gain ratio method where the difference in the score of the best and second best quantitative variables (R6 and R8 respectively) is much more than the difference between the second and 5th best quantitative variable (R1).

Summed ranking system: In this system, under each selection method, each independent quantitative and qualitative variable was ranked according to the score assigned by each technique. The rankings of individual variables were subsequently summed up across the 13 techniques. Positions were then assigned to each variable based on the summed ranking (see Table 8.4). In this system, the smaller the summed ranking value, the higher the position of the variable since, for example, a value of one (i.e. first) is better than five (i.e. fifth). This system gives all selection techniques equal chances of contribution to the final rank of each variable.

Final ranking system: This system was simply based on the average value of the percentage score ranking and summed ranking. A re-ranking of the variables was carried out based on this average values (see Table 8.5). Like in the case of the summed ranking system, the smaller the average rank value, the better the new position of the variable. Using this final ranking system, the top 7 quantitative and qualitative variables were selected to develop the models. These are presented along with the dependent variable in Table 8.6.

The dependent variable is named 'status', referring to the status of the construction firms (i.e. failed or existing). It is a binary variable with a value of one for existing firms and a value of zero for failed firms. A threshold of 0.5, which is the default threshold variable selection and model building, was used to separate the classes. In essence, a construction firm with a predicted value of over 0.5 is classified (predicted) as existing/healthy while one with a predicted value of below 0.5 is classified (predicted) as failed/failing**.**

Table 8.2: The raw scores assigned to the variables by each selection method

Variables	**Information gain**	**Kruskal test**	**mRMR**	**Chi-squared**	**Gain ratio**	**ANOVA test**	**Cforest importance**	**oneR**	**Relief**	**RF importance**	**SU**
Quantitative											
R1	0.18	164.34	0.09	0.56	0.18	152.53	0.02	1.73	0.02	0.02	0.21
R2	0.19	165.83	0.03	0.57	0.18	149.13	0.02	1.74	0.02	0.03	0.22
R3	0.1	108.27	0.04	0.44	0.15	88.99	0.01	1.89	0.01	0.01	0.15
R4	0.11	81.95	0	0.43	0.13	95.73	0.01	1.75	0.04	0.01	0.14
R5	0.12	69.84	0.15	0.45	0.14	60.8	0.01	1.75	0.04	0.01	0.15
R6	0.38	341.11	0.01	0.8	0.34	135.37	0.04	1.9	0.08	0.12	0.42
R7	0.09	48.3	-0.05	0.4	0.09	34.27	0	1.41	0.02	0.01	0.1
R8	0.37	157.6	-0.01	0.76	0.25	223.79	0.06	1.45	0.07	0.07	0.34
R9	0.1	40.49	0.13	0.42	0.09	28.85	0.01	1.45	0.05	0.01	0.11
R10	0.28	281.14	0	0.69	0.2	340.27	0.06	1.7	0.12	0.05	0.27
R11	0.05	0.76	0.04	0.29	0.08	1.88	0.01	1.69	0.04	0.01	0.08
Qualitative											
Q1	0.24	13.76	0	0.62	0.11	6.15	0	0.88	0.11	0.02	0.17

Variables	Information gain	Kruskal test	mRMR	Chi-squared	Gain ratio	ANOVA test	Cforest importance	oneR	Relief	RF importance	SU
Q2	0.22	27.11	-0.03	0.62	0.13	36.92	0	1.27	0.1	0.01	0.18
Q3	0.22	22.5	0.02	0.61	0.11	40.61	0.01	0.47	0.11	0.01	0.16
Q4	0.28	102.82	0.01	0.67	0.17	132.04	0.04	1.33	0.07	0.06	0.24
Q5	0.14	8.93	0.15	0.48	0.18	16.07	0.01	1.91	0.04	0.02	0.19
Q6	0.08	5.45	-0.01	0.37	0.08	0.95	0.01	1.52	0.07	0.01	0.09
Q7	0.09	39.98	0.02	0.4	0.21	18.55	0	1.96	0.08	0.02	0.16
Q8	0.23	140.78	0	0.63	0.16	131.14	0.04	1.2	0.12	0.06	0.22
Q9	0.18	26.25	0.02	0.54	0.1	13.96	0	0.82	0.05	0.01	0.14
Q10	0.05	57.97	0.11	0.31	0.07	50.3	0.01	1.75	0.09	0.01	0.07
Q11	0.04	19.25	0.09	0.27	0.07	29.09	0	1.69	0.05	0.01	0.06
Q12	0.12	25.2		0.46	0.1	18.09	0.01	1.13	0.12	0.02	0.13
Q13	0.12	101.74	0.03	0.46	0.12	87.6	0.02	1.72	0.04	0.03	0.14

MRMR: Minimum redundancy, maximum relevance filter

SU: Symmetrical Uncertainty

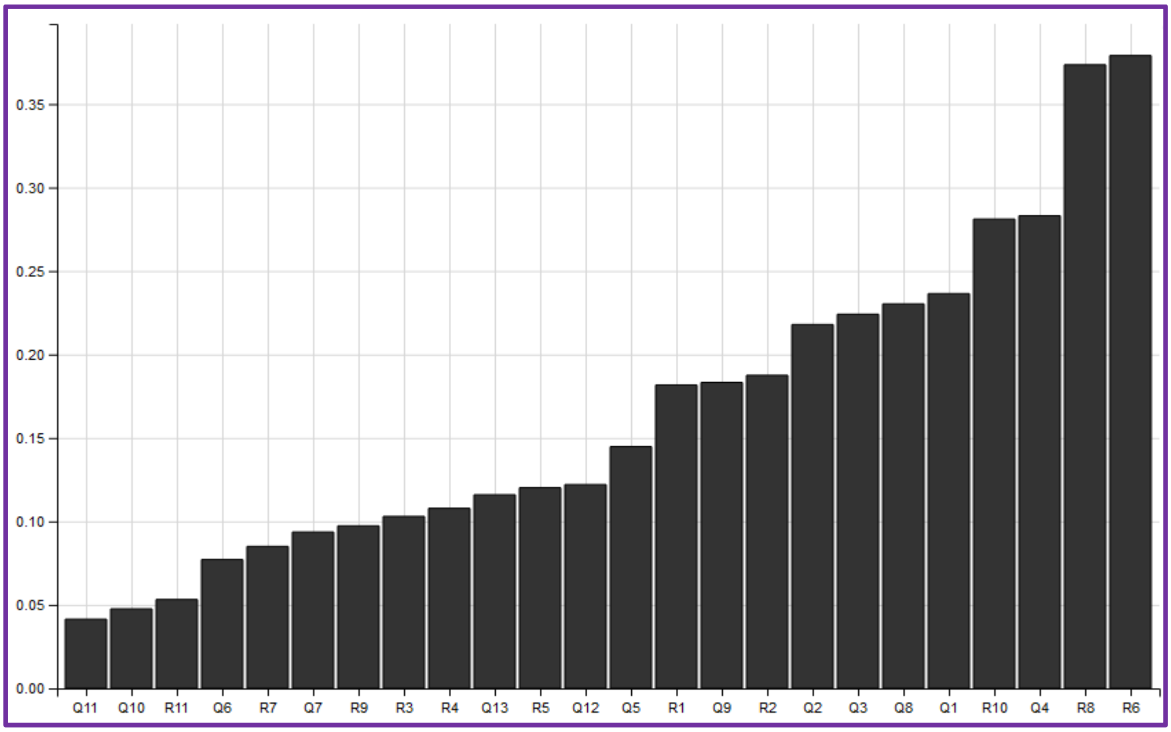


Figure 8.1: Chart of scores assigned by the Information Gain selection method

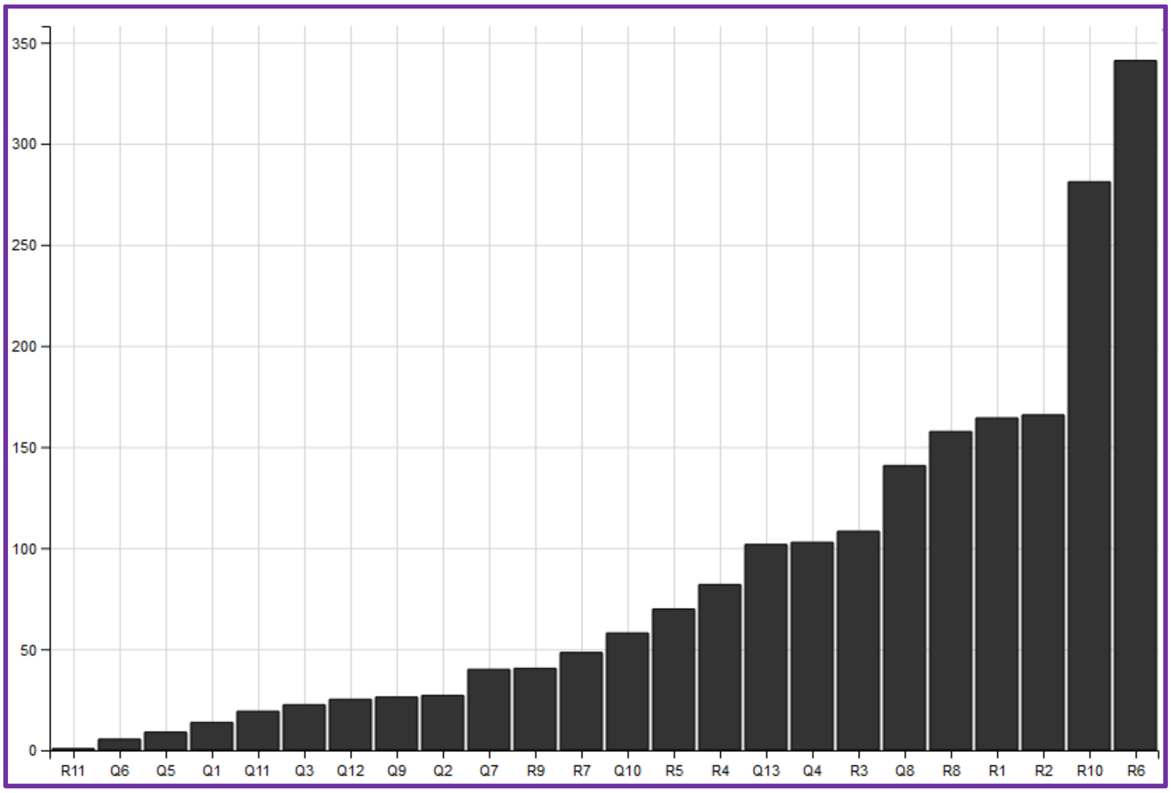


Figure 8.2: Chart of scores assigned by the Kruskal Test selection method

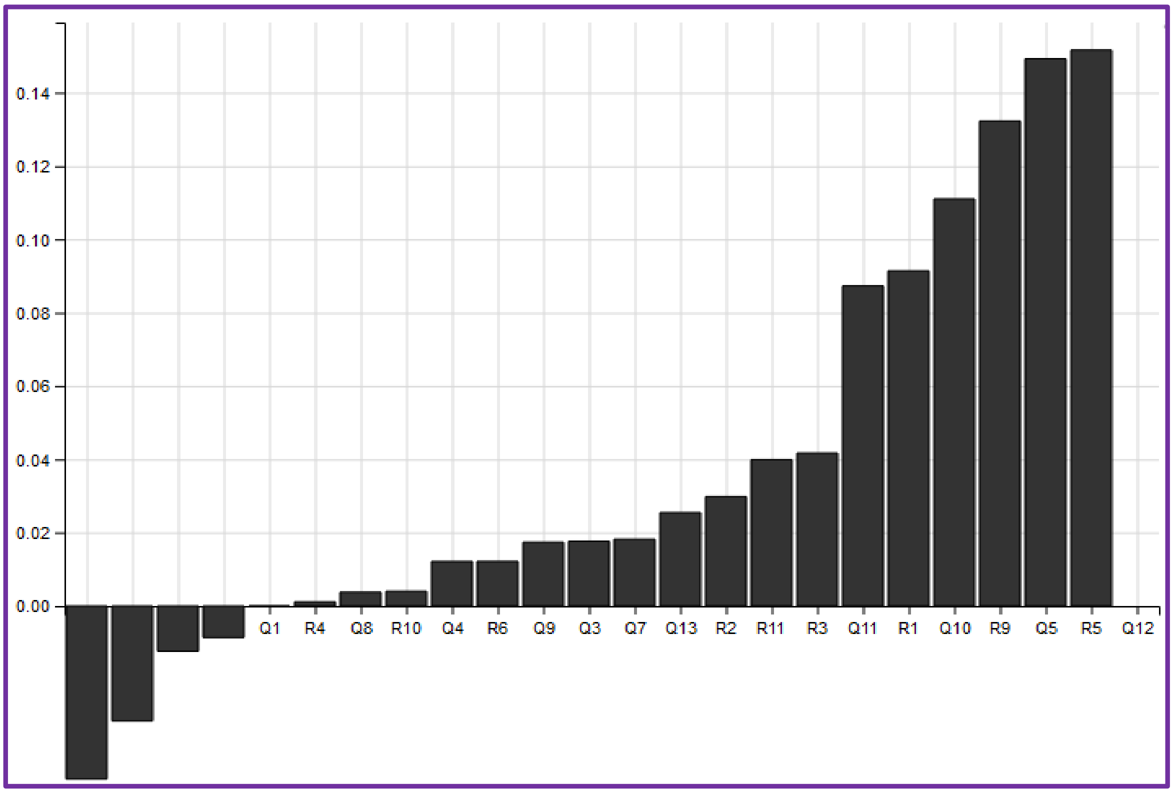


Figure 8.3: Chart of scores assigned by the mRMR selection method

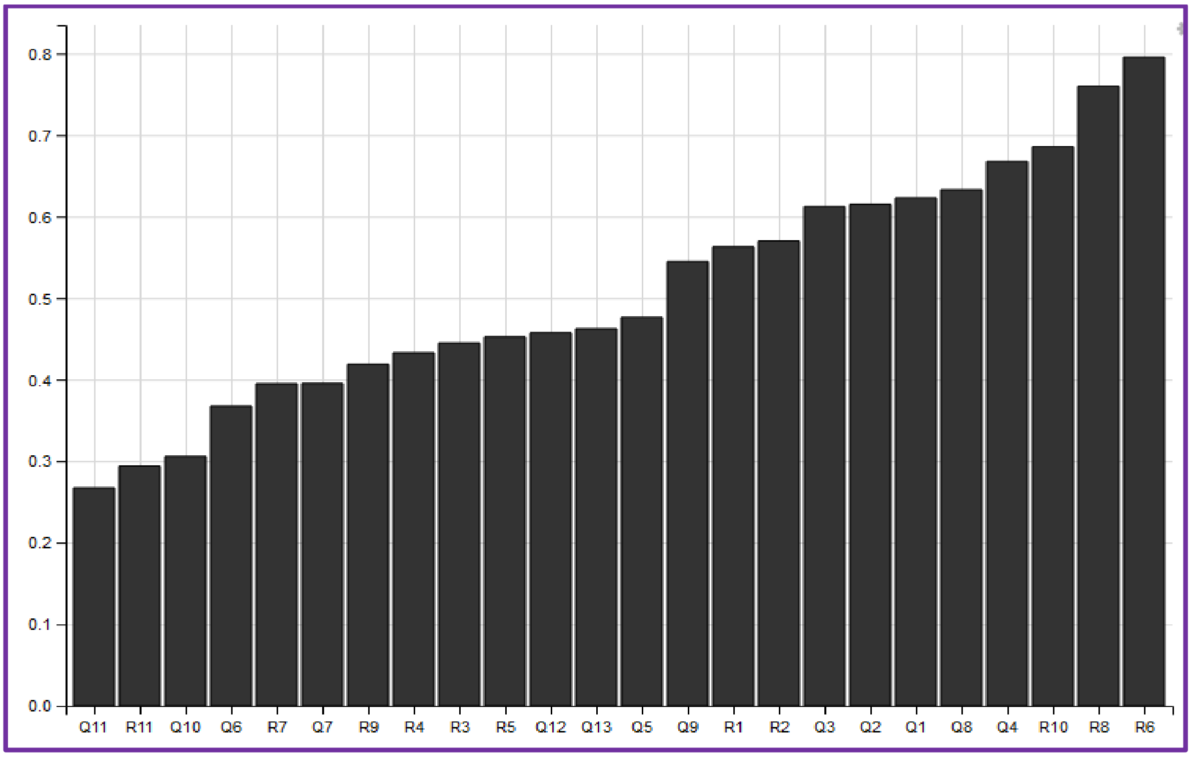


Figure 8.4: Chart of scores assigned by the Chi-Squared selection method

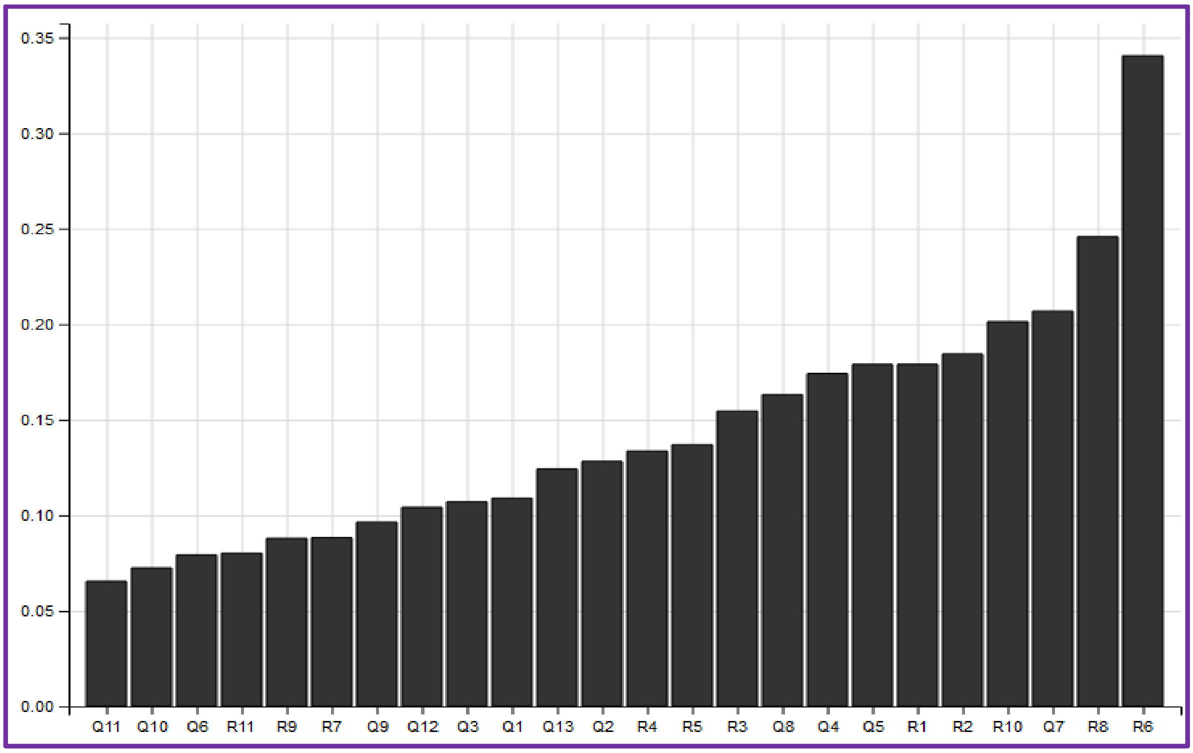


Figure 8.5: Chart of scores assigned by the Gain Ratio selection method

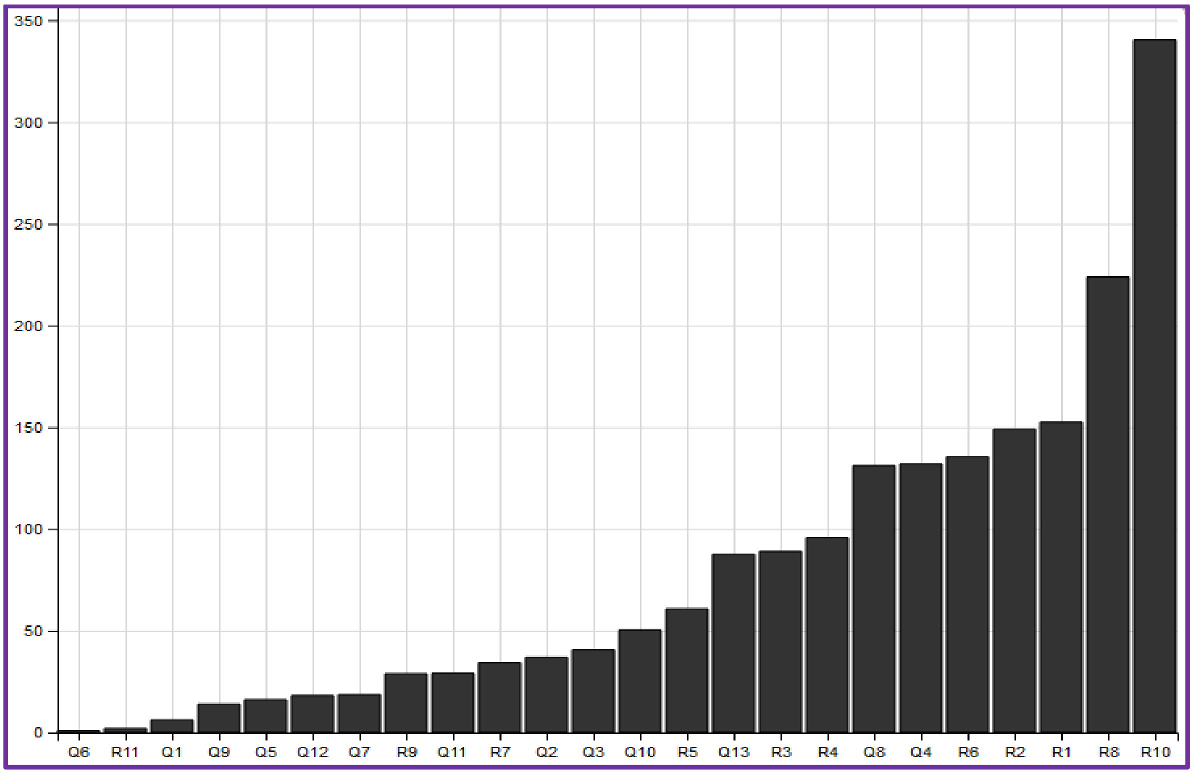


Figure 8.6: Chart of scores assigned by the ANOVA Test selection method

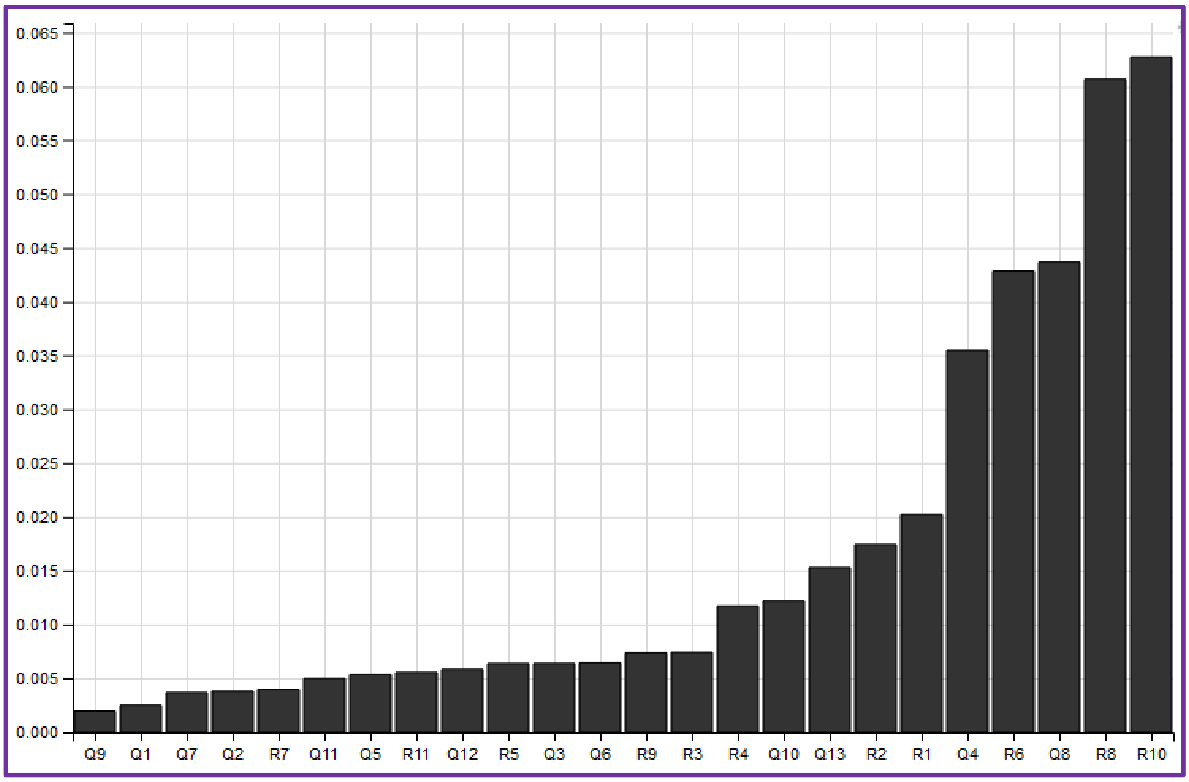


Figure 8.7: Chart of scores assigned by the CForest Importance selection method

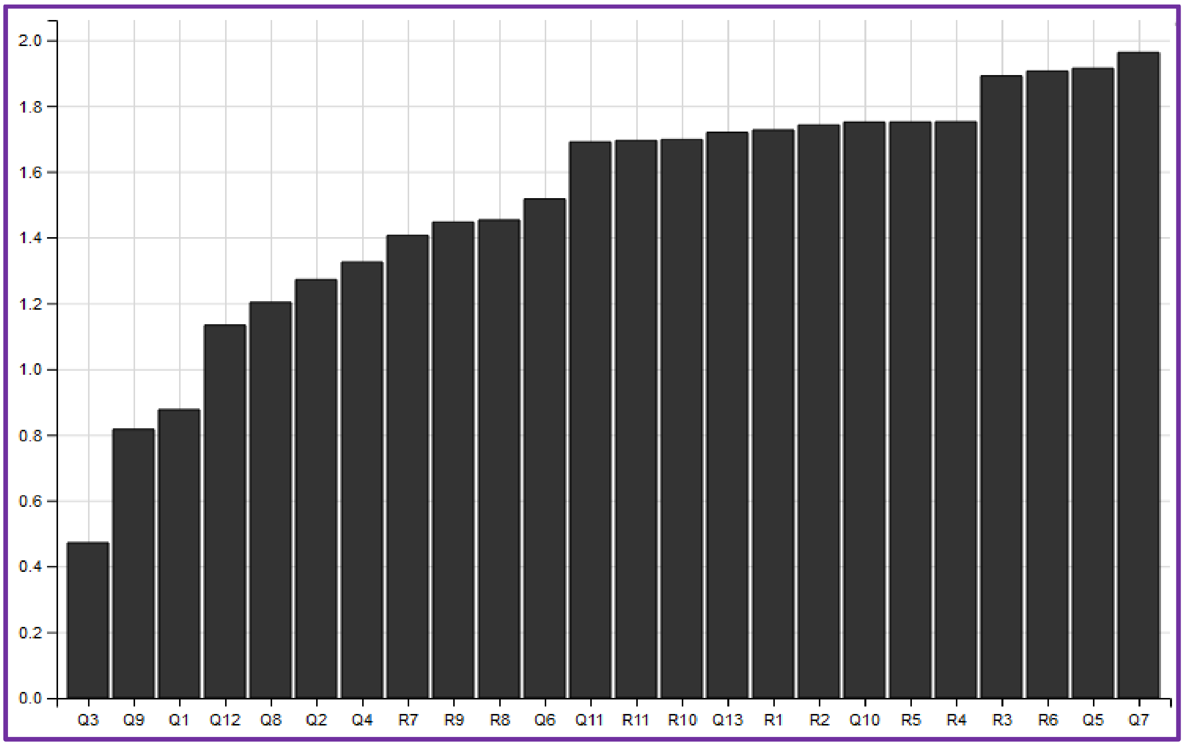


Figure 8.8: Chart of scores assigned by the oneR selection method

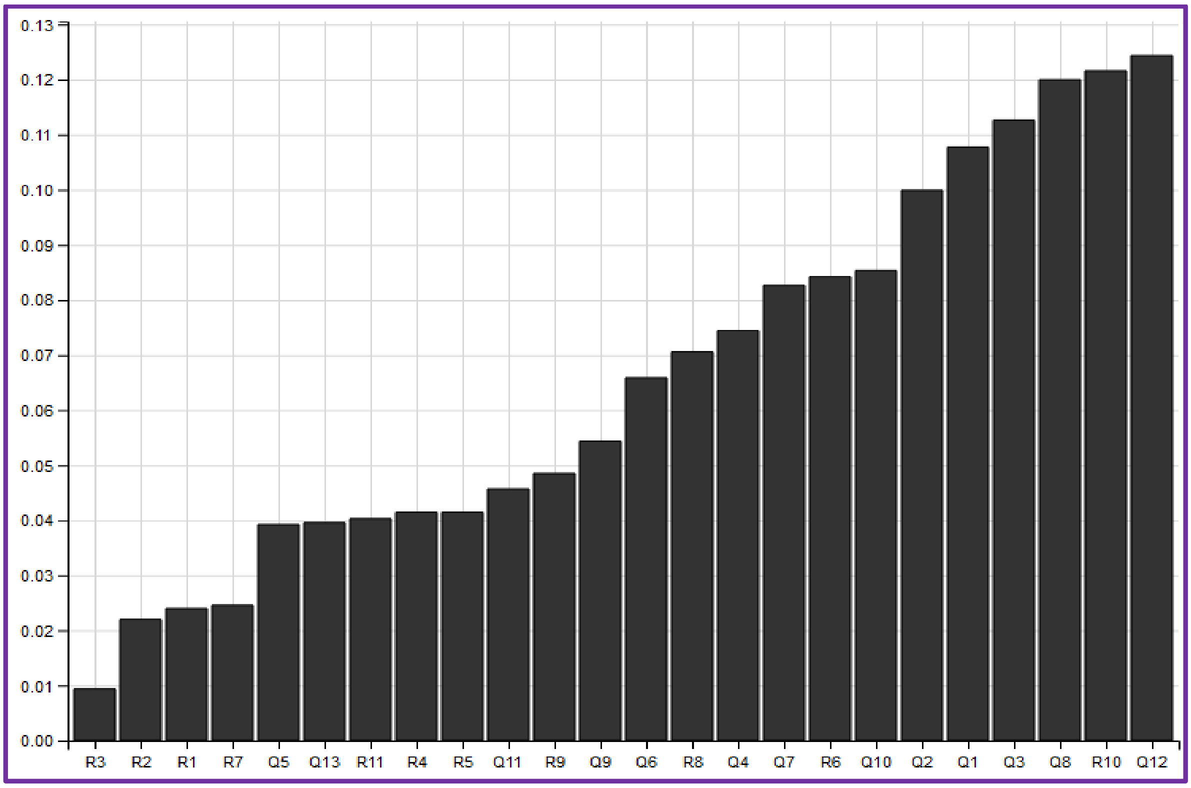


Figure 8.9: Chart of scores assigned by the Relief selection method

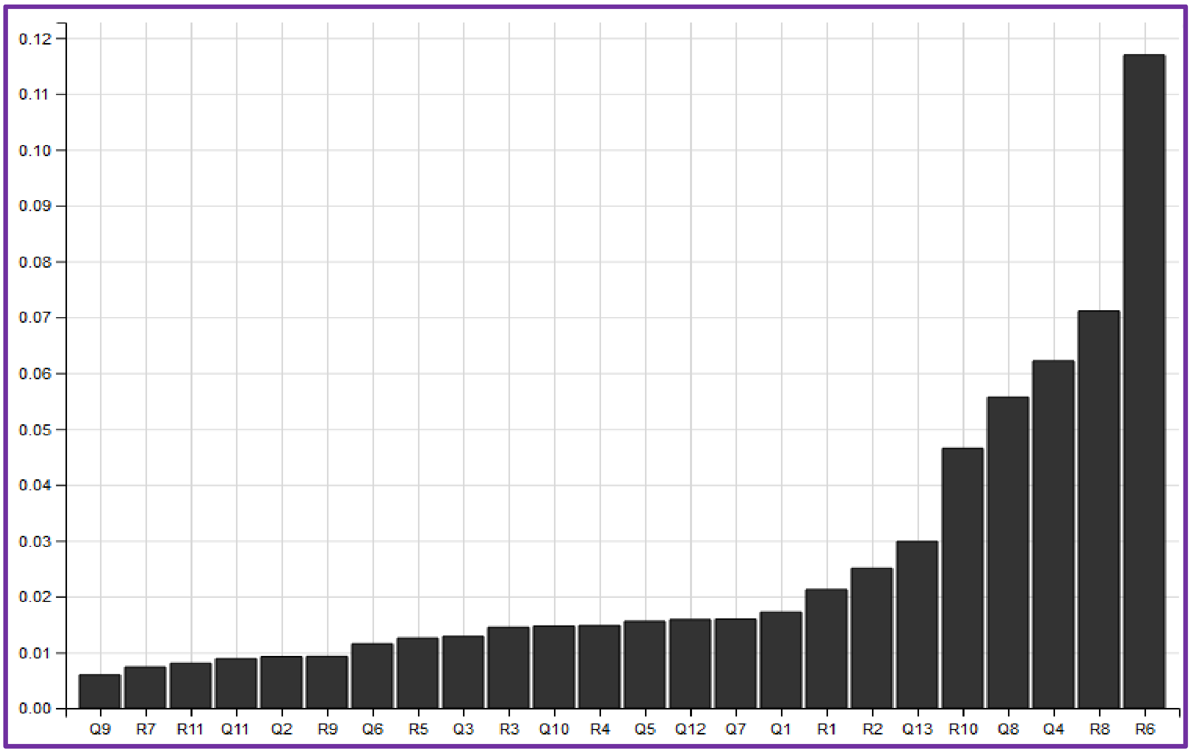


Figure 8.10: Chart of scores assigned by the Rf importance selection method

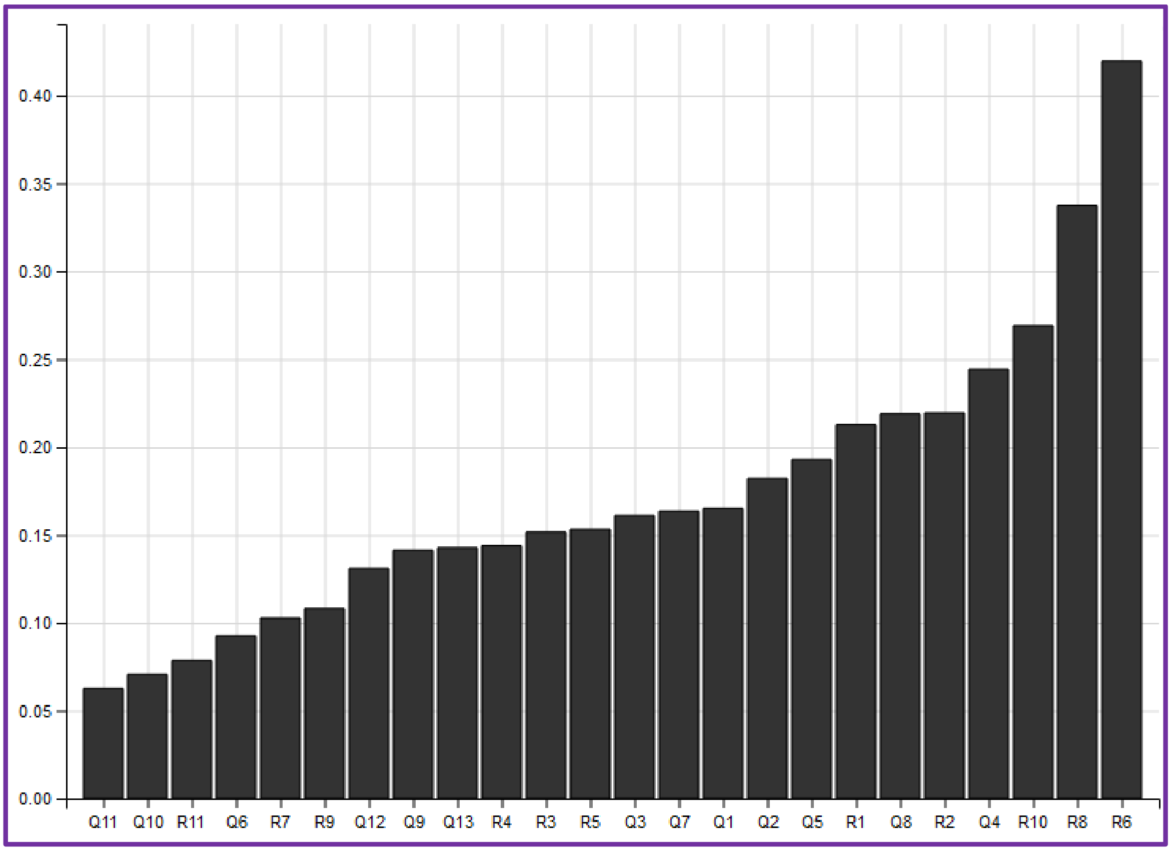


Figure 8.11: Chart of scores assigned by the Symmetrical uncertainty selection method

Table 8.3: Ranking based on Percentage score of quantitative and qualitative variables under each selection method

Variables	**Percentage score of variable under each selection method**											**Total Percentage**	**Ranking**
	Information gain	**Kruskal test**	**mRMR**	**Chi-squared**	**Gain ratio**	**ANOVA test**	**Cforest importance**	**oneR**	**Relief**	**RF importance**	**SU**		
Quantitative													
R1	4.52	8.01	10.71	4.57	5.23	8.06	5.00	4.79	1.28	3.13	5.07	60.38	**4th**
R2	4.77	8.08	3.57	4.65	5.23	7.88	5.00	4.82	1.28	4.69	5.31	55.29	**5th**
R3	2.51	5.28	4.76	3.59	4.36	4.70	2.50	5.23	0.64	1.56	3.62	38.77	**8th**
R4	2.76	3.99	0.00	3.51	3.78	5.06	2.50	4.85	2.56	1.56	3.38	33.96	**9th**
R5	3.02	3.40	17.86	3.67	4.07	3.21	2.50	4.85	2.56	1.56	3.62	50.33	**6th**
R6	9.55	16.63	1.19	6.53	9.88	7.15	10.00	5.26	5.13	18.75	10.14	100.22	**1st**
R7	2.26	2.35	-5.95	3.27	2.62	1.81	0.00	3.90	1.28	1.56	2.42	15.52	**11th**
R8	9.30	7.68	-1.19	6.20	7.27	11.82	15.00	4.02	4.49	10.94	8.21	83.73	**3rd**
R9	2.51	1.97	15.48	3.43	2.62	1.52	2.50	4.02	3.21	1.56	2.66	41.47	**7th**
R10	7.04	13.70	0.00	5.63	5.81	17.97	15.00	4.71	7.69	7.81	6.52	91.90	**2nd**
R11	1.26	0.04	4.76	2.37	2.33	0.10	2.50	4.68	2.56	1.56	1.93	24.09	**10th**
Qualitative													
Q1	6.03	0.67	0.00	5.06	3.20	0.32	0.00	2.44	7.05	3.13	4.11	32.00	**8th**
Q2	5.53	1.32	-3.57	5.06	3.78	1.95	0.00	3.52	6.41	1.56	4.35	29.91	**10th**
Q3	5.53	1.10	2.38	4.98	3.20	2.15	2.50	1.30	7.05	1.56	3.86	35.61	**6th**
Q4	7.04	5.01	1.19	5.47	4.94	6.97	10.00	3.68	4.49	9.38	5.80	63.97	**2nd**
Q5	3.52	0.44	17.86	3.92	5.23	0.85	2.50	5.29	2.56	3.13	4.59	49.88	**3rd**

Variables	**Percentage score of variable under each selection method**											**Total Percentage**	**Ranking**
	Information gain	**Kruskal test**	**mRMR**	**Chi-squared**	**Gain ratio**	**ANOVA test**	**Cforest importance**	**oneR**	**Relief**	**RF importance**	**SU**		
Q6	2.01	0.27	-1.19	3.02	2.33	0.05	2.50	4.21	4.49	1.56	2.17	21.41	**13th**
Q7	2.26	1.95	2.38	3.27	6.10	0.98	0.00	5.43	5.13	3.13	3.86	34.49	**7th**
Q8	5.78	6.86	0.00	5.14	4.65	6.93	10.00	3.32	7.69	9.38	5.31	65.07	**1st**
Q9	4.52	1.28	2.38	4.41	2.91	0.74	0.00	2.27	3.21	1.56	3.38	26.66	**12th**
Q10	1.26	2.83	13.10	2.53	2.03	2.66	2.50	4.85	5.77	1.56	1.69	40.77	**5th**
Q11	1.01	0.94	10.71	2.20	2.03	1.54	0.00	4.68	3.21	1.56	1.45	29.33	**11th**
Q12	3.02	1.23	0.00	3.76	2.91	0.96	2.50	3.13	7.69	3.13	3.14	31.45	**9th**
Q13	3.02	4.96	3.57	3.76	3.49	4.63	5.00	4.76	2.56	4.69	3.38	43.81	**4th**

Table 8.4: Ranking based on summation of ranking of quantitative and qualitative variables under each selection method

Variables	Ranking of variable according to score under each selection method											Sum of all positions	Summed Ranking
	Information gain	Kruskal test	mRMR	Chi-squared	Gain ratio	ANOVA test	Cforest importance	oneR	Relief	RF importance	SU		
Quantitative													
R1	5	4	3	5	4	3	4	6	8	5	5	52	**5th**
R2	4	3	6	4	4	4	4	5	8	4	4	50	**4th**
R3	8	6	4	7	6	7	6	2	11	6	6	69	**7th**
R4	7	7	8	8	8	6	6	3	5	6	8	72	**8th**
R5	6	8	1	6	7	8	6	3	5	6	6	62	**6th**
R6	1	1	7	1	1	5	3	1	2	1	1	24	**1st**
R7	10	9	11	10	9	9	11	11	8	6	10	104	**11th**
R8	2	5	10	2	2	2	1	9	3	2	2	40	**3rd**
R9	8	10	2	9	9	10	6	9	4	6	9	82	**9th**
R10	3	2	8	3	3	1	1	7	1	3	3	35	**2nd**
R11	11	11	4	11	11	11	6	8	5	6	11	95	**10th**
Qualitative													
Q1	2	11	9	3	7	12	9	11	3	4	5	76	**8th**
Q2	4	6	13	3	5	6	9	8	5	8	4	71	**7th**
Q3	4	9	5	5	7	5	4	13	3	8	6	69	**6th**

Variables	**Ranking of variable according to score under each selection method**											**Sum of all positions**	**Summed Ranking**
	Information gain	**Kruskal test**	**mRMR**	**Chi-squared**	**Gain ratio**	**ANOVA test**	**Cforest importance**	**oneR**	**Relief**	**RF importance**	**SU**		
Q4	1	2	8	1	3	1	1	7	8	1	1	34	**1st**
Q5	7	12	1	7	2	10	4	2	12	4	3	64	**4th**
Q6	11	13	12	11	11	13	4	6	8	8	11	108	**13th**
Q7	10	5	5	10	1	8	9	1	7	4	6	66	**5th**
Q8	3	1	9	2	4	2	1	9	1	1	2	35	**2nd**
Q9	6	7	5	6	9	11	9	12	10	8	8	91	**11th**
Q10	12	4	2	12	12	4	4	3	6	8	12	79	**9th**
Q11	13	10	3	13	12	7	9	5	10	8	13	103	**12th**
Q12	8	8	9	8	9	9	4	10	1	4	10	80	**10th**
Q13	8	3	4	8	6	3	3	4	12	3	8	62	**3rd**

Table 8.5: Ranking based on average values from percentage score and summed ranking systems

Variables	Percentage score ranking	Summed ranking	Average of percentage score and summed ranking	Ranking based on average values
Quantitative				
R1	4	5	4.5	**4th**
R2	5	4	4.5	**4th**
R3	8	7	7.5	**7th**
R4	9	8	8.5	**9th**
R5	6	6	6	**6th**
R6	1	1	1	**1st**
R7	11	11	11	**11th**
R8	3	3	3	**3rd**
R9	7	9	8	**8th**
R10	2	2	2	**2nd**
R11	10	10	10	**10th**
Qualitative				
Q1	8	8	8	**8th**
Q2	10	7	8.5	**9th**
Q3	6	6	6	**5th**
Q4	2	1	1.5	**1st**
Q5	3	4	3.5	**3rd**
Q6	13	13	13	**13th**
Q7	7	5	6	**5th**
Q8	1	2	1.5	**1st**
Q9	12	11	11.5	**11th**
Q10	5	9	7	**7th**
Q11	11	12	11.5	**11th**
Q12	9	10	9.5	**10th**
Q13	4	3	3.5	**3rd**

Table 8.6: Details of the top seven ranked quantitative and qualitative independent variables selected to develop the model and the dependent variable

Variable category	Serial number	Variable name	Assigned (quantitative) variable code
Quantitative variables	1.	Solvency ratio (Asset-based) (%)	R6
	2.	Working Capital per employee (unit)	R10
	3.	Average Remuneration per employee (unit)	R8
	4.	Return on Shareholders' Funds (%)	R1
	5.	Return on Capital Employed (%)	R2

Variable category	Serial number	Variable name	Assigned (quantitative) variable code
Quantitative variables	6.	Liquidity ratio	R5
	7.	Return on Total Assets (%)	R3
Quantitative variables	1.	Strategic and external issues	Q4
	2.	Finance and conflict related issues	Q8
	3.	Performance on projects	Q5
	4.	Industry contract/project knowledge	Q13
	5.	Project related (external) conflict	Q7
	6.	Top management characteristics 1	Q3
	7.	Profit issues	Q10
Dependent Variable		Status (failed or existing)	

8.5 Model development

8.5.1 Data pre-processing

To develop the model, the default working directory of the ‘R’ programming language software was set to the ‘Spot Instances’ on the AWS EC2 while the executioner in ‘R’ was configured to operate with the Spark Master on the NameNode. The full data on the Hadoop DataNodes was subsequently assigned a new name: ‘FacLogData1’.

Since the data was arranged in such a way that those of the failed construction firms were grouped together on one side and those of the existing on the other, there was a potential problem of the prediction tools recognising this pattern and making predictions based on the arrangement. The first step was thus to randomise the arrangement of the data in such a way that data for failed and existing construction firms are interwoven in no specific pattern. The ‘runif’ command was used to execute this process.

Next step was to install ‘sparklyr’ which is the Spark executor on ‘R’. The sparklyr was subsequently connected to the Spark Master on the NameNode on the AWS EC2 using the spark_connect function to connect the Instance address which is CF-IPM.BigDataCluster (see section 8.2). The ‘mlr’ machine learning executor with CRAN on ‘R’ was installed to be used with Spark. Subsequently the ‘makeClassifTask’ function was used to create the task on the data using the 70% of the data required for training and specifying the target (dependent) variable as ‘status’. The task was named ‘task’ and was used to create all the models using the ‘mlr’ executor with the integrated tools/algorithms. The packages and libraries of the integrated tools/algorithms were installed and unpacked. The code for these processes, including packages for some of the tools/algorithms used, are given.

```
>gp = runif(nrow(FacLogData1))
>FacLogData1 = FacLogData1[order(gp),]
>str(FacLogData1)
>summary(FacLogData1)
>SP = spark_connect(master = "spark://local: CF-IPM.BigData
Cluster")
>task = makeClassifTask(data = Train, target = "Status")
>install.packages("sparklyr")
>install.packages("mlr")
>install.packages("randomForestSRC")
>install.packages("imp.learner ")
>install.packages("kernlab ")
>install.packages("caTools ")
>install.packages("GA")
>install.packages("ada")
>install.packages("kohonen ")
>install.packages("stats")
>install.packages("nnet")
>Libraray(sparklyr)
>Libraray(mlr)
>library(imp.learner)
>library(randomForestSRC)
>library(kernlab)
>library(caTools)
>library(GA)
>library(ada)
>library(stats)
>library(nnet)
```

8.5.2 The prediction tools/algorithms used to build the models

Thirteen tools/algorithms, including very powerful unpopular artificial intelligence tools, were used to develop more than 13 construction firms insolvency prediction models (CF-IPM). Some of the popular ones, including multiple discriminant analysis, logistic regression, support vector machines, K-nearest neighbour and adaptive boosting, are explained.

Multiple Discriminant Analysis (MDA): Though MDA can be linear (LDA) or quadratic (QDA), the LDA is much simpler and much more popular (Balcaen and Ooghe 2006) hence it will be the one explained here. The MDA model uses a linear combination of variables, which are normally financial ratios, which best differentiate between failing and surviving firms to classify firms into one of the two groups. The first MDA model, called Z-score or Z model, was developed by Altman (1968) and was first applied to the construction industry by Mason and Harris (1979). The MDA function, constructed after variable selection, is as follows (Altman 1968):

$Z = c_1X1 + c_2X2 + \ldots\ldots\ldots\ldots\ldots\ldots.. + c_nXn$

Where $c_1, c_2, \ldots\ldots\ldots\ldots\ldots\ldots.. c_n,$ = discriminant coefficients

And X1, X2, …………….. Xn = independent variables

The LDA calculates the discriminant coefficients, cj, while Xj represents values of the selected variables, Where j = 1,2, …………. N. The function can be used to calculate the Z-score of any firm which possesses the independent variables in the function. A cut-off Z-score is chosen based on the scores of failing and surviving sample firms and used to classify newly assessed firms.

Logistic regression (LR): LR is a "conditional probability model which uses the non-linear maximum log-likelihood technique to estimate the probability of firm failure under the assumption of a logistic distribution" (Jackson and Wood, 2013, p. 190). It was first applied to bankruptcy prediction by (Ohlson, 1980). Based on (Hosmer, Lemeshow and Sturdivant, 2013), the LA function, constructed after variable selection, is as follows (Ohlson, 1980):

$P_1(V_i) = 1/[1 + \exp - (b_0 + b_1V_{i1} + b_2V_{i2} + b_2V_{i2} + \ldots\ldots + b_nV_{in})] = 1/[1 + \exp - (D_i)]$

where $P_1(V_i)$ = probability of failure given the vector of attributes V_i;

V_{ij} = value of attribute or variable j (j = 1, 2, ….., n) for firm *i*;

b_j = coefficient for attribute *j*; b_0 = intercept D_i = logit of firm i.

Artificial Neural Network (ANN): ANN was created to imitate how the neural system of the human brain works and was first applied to insolvency prediction by Odom and Sharda (1990). A typical ANN is a network of nodes interconnected in layers. There are some parameters, architectures, algorithms, and learning/training methods that can be used to develop an ANN (Jo, Han and Lee, 1997) and choosing the best combination can be demanding. The architecture of the ANN refers to the number of layers, number of nodes in each layer and the method of interconnectivity between nodes. Common architectures for ANNs include Multi-layer perceptron (MLP), self-organizing mapping and Perceptron (Chung, Tan and Holdsworth, 2008).

Support vector machines (SVM): SVM employs a linear model to develop an optimal separating hyperplane by using a highly non-linear mapping of input vectors into a high-dimensional feature space (Ravi Kumar and Ravi 2007). It constructs the boundary using binary class. The variables closest to the hyperplane are called support vectors and are used to define the binary outcome (failing or non-failing) of assessed firms. All other samples are ignored and are not involved in deciding the binary class boundaries (Vapnik, 1998). SVM is very simple because its mathematical analysis is easy to execute(Ravi Kumar and Ravi 2007). Like ANN, it has some parameters that have to be varied to perform optimally

K-nearest neighbour (KNN): This is a very straightforward algorithm that classifies a new sample (i.e. construction firm) based on the properties of the nearest neighbours. Where the properties of the neighbours differ widely (e.g. properties of failed and existing construction firms), the new sample is classified based on the majority class neighbours (Cunningham and Delany 2007). The distance away from the new sample within which other samples can be classified as nearest neighbours can be decided by the programmer.

Random forest (RF): This is an ensemble classifier which uses the bagging method. Ensemble classifiers use a particular algorithm repeatedly with various settings, which create a bootstrap sample of the data set, to yield different results (Liaw and Wiener 2002). A majority vote is then used for the final prediction. Random forest uses the decision tree algorithms (e.g. C4.5, CART, among others) for its ensemble.

Adaptive Boosting (AB): This is an ensemble classifier which uses the boosting method and like random forest, uses the decision tree algorithms for its ensemble. For every new tree in this ensemble, the classifier assigns extra importance to wrongly predicted objects in prior trees (Schapire *et al.*, 1998).

8.5.3 Understanding the features and outputs of the models

The tools/algorithms used have various levels of transparency. With some, a visible model (not necessarily interpretable) is created during development while with others, there is no visible model. In this subsection, the details that are later presented with the developed CF-IPMs are explained

Model: The model developed (on the training data) if available, usually appears in the form of equation (e.g. with LDA or LR) or a network diagram (e.g. with ANN). In the case of an equation, it is usually the case that the greater the coefficient assigned to an independent variable, the more important the variable is. Not all equation models are interpretable in this manner, however. With networks, networks of algorithms like decision trees are interpretable while those of ANN are not.

The model processing: This is the output given by 'R' when running the model. It applies to very few tools/algorithms

Confusion matrix: Each model will be presented alongside its confusion matrix. A confusion matrix is a 2 x 2 matrix (see Table 8.7) that displays the simplest form of result that emerges from the model testing/validation process. The result is based on the test data. Very important parameters like overall accuracy, sensitivity (or true positive rate), specificity (or true negative rate), Type I error (or false positive error rate) and Type II error (or false negative error rate) can be calculated from the confusion matrix

Table 8.7: A Standard confusion matrix result for a model (Torgo, 2011)

	Predicted class (failed firm) = 0	**Predicted class (existing firm) = 1**
Actual class (failed firm) = 0	True Negatives (TN)	False Positives (FP)
Actual class (existing firm) = 1	False Negatives (FN)	True Positives (TP)

Overall accuracy: This is the ratio of the total number of correctly predicted classes to the total number of sample construction firms in the test data. The equation used to calculate overall accuracy from the confusion matrix is given here. The ‘N’ in the equation represent total number of sample construction firms in the test data

$$Overall\ accuray = \frac{TN + TP}{N}\ \text{(Torgo 2011)}$$

Sensitivity: This is also known as true positive rate. It is the ratio of existing construction firms correctly predicted as existing to the total number of existing construction firms in the test data. It reveals the percentage of the actual existing construction firms predicted correctly by a model. The equation used to calculate sensitivity from the confusion matrix is:

$$\text{Sensitivity} = \frac{TP}{TP + FN}\text{(Torgo 2011)}$$

Specificity: This is also known as true negative rate. It is the ratio of failed construction firms correctly predicted as failed to the total number of failed construction firms in the test data. It reveals the percentage of the actual failed construction firms predicted correctly by a model. The equation used to calculate specificity from the confusion matrix is:

$$\text{Specificity} = \frac{TN}{TN + FP}\ \text{(Torgo 2011)}$$

Type I error: This is also known as false positive error rate. It is the ratio of failed construction firms wrongly predicted as existing to the total number of failed construction firms in the test data. It reveals the percentage of the actual failed construction firms predicted wrongly by a model. Type I error is costlier than Type II error. The equation used to calculate Type I error from the confusion matrix is:

$$\text{Type I error} = \frac{FP}{TN + FP}\ \text{(Torgo 2011)}$$

Type II error: This is also known as false negative error rate. It is the ratio of existing construction firms wrongly predicted as failed to the total number of existing construction firms in the test data. It reveals the percentage of the actual existing construction firms predicted wrongly by a model. The equation used to calculate Type II error from the confusion matrix is:

$$\text{Type II error} = \frac{FN}{TP + FN}\ \text{(Torgo 2011)}$$

Receiver operator characteristic (ROC) curve: Each model will be presented alongside its ROC curve which is a plot of sensitivity (true positive rate) on the y-axis against specificity (false positive rate) on the x-axis. The threshold of the model, which varies from zero to one with a scale of 0.1, is normally shown on the vertical axis to the right of the plot and sometimes on the curve as well. The plot can be used to identify the best threshold, usually between 0.1 and 0.9, as against just adopting the default 0.5 threshold. A better threshold can increase the accuracy of a model based on reduced costlier error type (i.e. Type I error in this case). The threshold in the proposed solution will not be changed from the default value to allow a fairer comparison between the models.

Area under the curve (AUC): This is the area under the ROC curve which is widely accepted as the best measure of the performance of a model. The AUC value of similar models could even be different, making it easy to pick the better model. Since the maximum value of specificity and sensitivity which make up the axes of ROC curve are one, then the maximum AUC value, which represents excellent accuracy, is one. A model with an AUC value below 0.5 has a less than average performance which is considered to be very poor.

8.5.4 Details of the development of the CF-IPMs

The task created with the 'makeClassifTask' function, using training data (i.e. 70% of the data), was used in the process of developing the models. The 'makeLearner' function combined with 'classif.algorithm' was used to create the necessary prediction algorithm, which is normally termed as the learner in mlr. The learner was trained with the task created using the 'train' function, thereby developing a CF-IPM. The CF-IPM was then tested by running it on the test data (i.e. the remaining 30% of the data) using the 'Task.pred' function. The confusion matrix of the result of the test, which provides the information that allows a fair comparison of the results, is then called using the 'getConfMatrix' function. This is followed by plotting the ROC curve using the 'plotROCCurves' and 'generateThreshVsPerfData' functions. Finally, the AUC value, which allows for an even more accurate models' performance comparison to the smallest of margins, is generated using the 'as.numeric' and 'convert, "auc"' functions.

Tables 8.8 to 8.20 in the model presentation in the following pages display the confusion matrices of the 13 Big Data CF-IPMs developed in the proposed solution while Figures 8.12 to 8.26, excluding Figures 8.18 and 8.20, display the ROC plots of the Big Data CF-IPMs.

Figures 8.18 and 8.20 are displays of the Big Data artificial neural network and decision tree models. Once the necessary codes were entered, the system developed the models in seconds because of the use of Big Data Analytics. As a simple process to check the effect of the Big Data Analytics platform, an ANN model which was highly tuned with 1e9 iterations was attempted on a regular computer. The model was not developed in over two days. The process was consequently aborted

8.5.4.1 The Big Data linear (multiple) discriminant analysis CF-IPM

The code of the Big Data Linear Discriminant Analysis (LDA) CF-IPM

```
LDA_lrn = makeLearner("classif.lda", predict.type = "prob")
LDA_mod = train(LDA_lrn, task)
LDA_task2 = makeClassifTask(data = Test, target = "Status")
LDA_Task.pred = predict(LDA_mod, LDA_task2, predict.type = "prob")
getConfMatrix(LDA_Task.pred)
LDA_ROCR_MLR =
plotROCCurves(LDA_ROCR_MLR)
LDA_ROCRconvert = asROCRPrediction(LDA_Task.pred)
LDA_ROCR_ROCRcalc   =   ROCR::performance(LDA_ROCRconvert,   "tpr",
"fpr")
LDA_auc = as.numeric(performance(LDA_ROCRconvert, "auc")@y.values)

LDA_auc
```

The LDA model: $131.40R6 - 1.07R10 + 1.16R8 + 123.35R1 + 6.73R2 - 0.23R5 - 130.54R3 - 1.84Q4 - 0.41Q8 - 0.18Q5 - 0.32Q13 + 0.27Q7 - 0.43Q3 - 0.27Q10$

Table 8.8: **The confusion matrix** *of the Big Data LDA CF-IPM*

	Predicted as failed firm (0)	**Predicted as existing firm (1)**
Actual failed firm (0)	116	39
Actual existing firm (1)	44	119

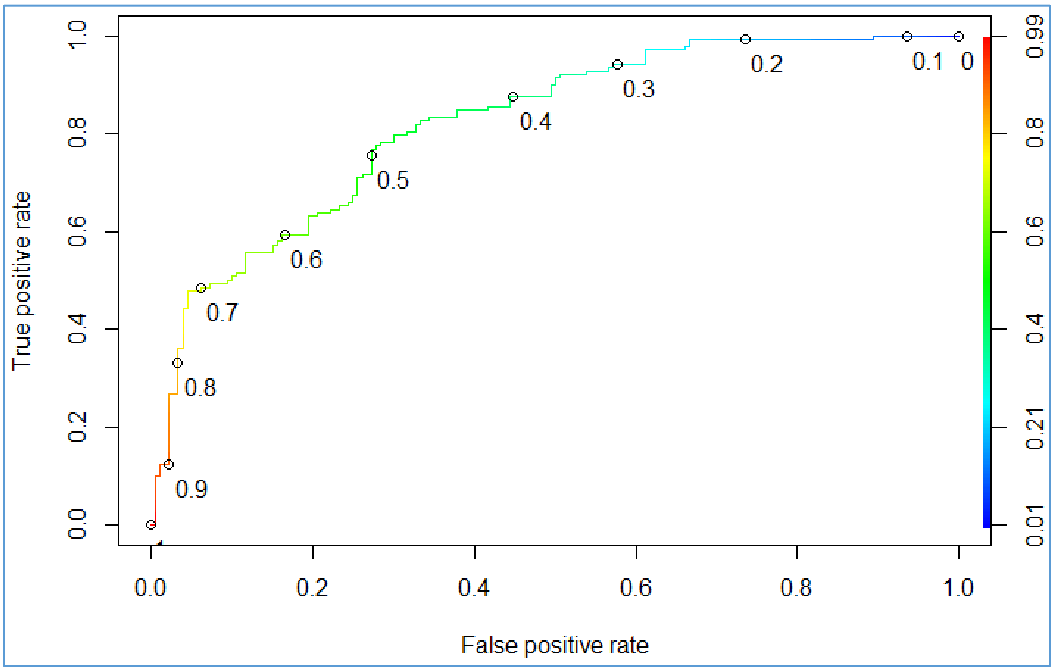


Figure 8.12: **The ROC curve** *of the Big Data LDA CF-IPM*

The AUC value of the Big Data LDA CF-IPM = 0.821095

8.5.4.2 The Big Data quadratic (multiple) discriminant analysis CF-IPM

The code of the Big Data Quadratic Discriminant Analysis (QDA) CF-IPM

```
QDA_lrn = makeLearner("classif.binomial", predict.type = "prob")
QDA_mod = train(QDA_lrn, task1)
QDA_task2 = makeClassifTask(data = Test1, target = "Status")
QDA_Task.pred = predict(QDA_mod, QDA_task2, predict.type = "prob")
getConfMatrix(QDA_Task.pred)
plotROCCurves(QDA_ROCR_MLR)
QDA_ROCRconvert = asROCRPrediction(QDA_Task.pred)
QDA_ROCR_ROCRcalc = ROCR::performance(QDA_ROCRconvert, "tpr",
"fpr")
QDA_auc = as.numeric(performance(QDA_ROCRconvert, "auc")@y.values)
QDA_auc
```

The QDA model: $\underline{-60.59 + 30.68R6 - 0.24R10 + 0.27R8 + 28.80R1 + 1.57R2 - 0.05R5 - 30.48R3 - 0.43Q4 + -0.09Q8 - 0.04Q5 - 0.07Q13 + 0.06Q7 - 0.10Q3 - 0.06Q10}$

Table 8.9: **The confusion matrix** *of the Big Data QDA CF-IPM*

	Predicted as failed firm (0)	**Predicted as existing firm (1)**
Actual failed firm (0)	110	45
Actual existing firm (1)	46	117

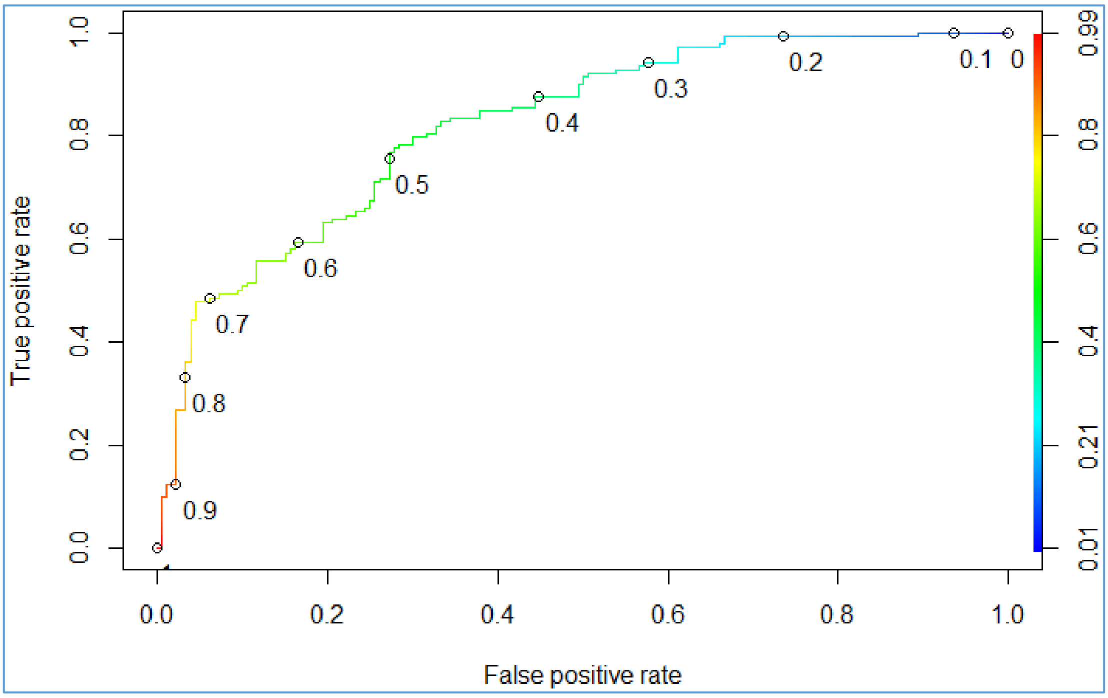


Figure 8.13: **The ROC curve** *of the Big Data QDA CF-IPM*

The AUC value of the Big Data QDA CF-IPM = 0.8210548

8.5.4.3 The big data logistic regression CF-IPM

The code of the Big Data Logistic Regression (LR) CF-IPM

```
LR_lrn = makeLearner("classif.logreg", predict.type = "prob")
LR_mod = train(LR_lrn, task1)
LR_task2 = makeClassifTask(data = Test1, target = "Status")
LR_Task.pred = predict(LR_mod, LR_task2, predict.type = "prob")
getConfMatrix(LR_Task.pred)
plotROCCurves(LR_ROCR_MLR)
LR_ROCRconvert = asROCRPrediction(LR_Task.pred)
LR_ROCR_ROCRcalc = ROCR::performance(LR_ROCRconvert, "tpr", "fpr")
LR_auc = as.numeric(performance(LR_ROCRconvert, "auc")@y.values)
LR_auc
```

The LR model: $\underline{-60.59 + 30.68R6 - 0.24R10 + 0.27R8 + 28.80R1 + 1.57R2}$
$\underline{- 0.05R5 - 30.48R3 - 0.43Q4 + -0.09Q8 - 0.04Q5 - 0.07Q13}$
$\underline{+ 0.06Q7 - 0.10Q3 - 0.06Q10}$

Table 8.10: **The confusion matrix** *of the Big Data LR CF-IPM*

	Predicted as failed firm (0)	**Predicted as existing firm (1)**
Actual failed firm (0)	110	45
Actual existing firm (1)	46	117

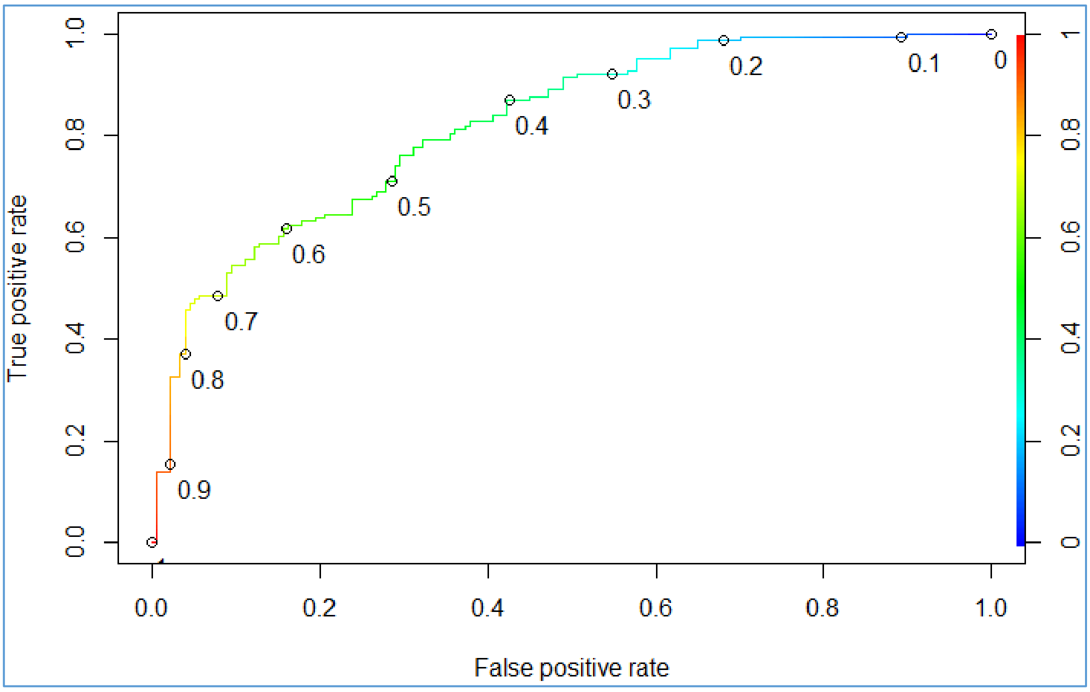


Figure 8.14: **The ROC curve** *of the Big Data LR CF-IPM*

The AUC value of the Big Data LR CF-IPM = 0.8210548

8.5.4.4 The big data naïve bayes CF-IPM

The code of the Big Data Naïve Bayes (NB) CF-IPM

```
NB_lrn = makeLearner("classif.naiveBayes", predict.type = "prob")
NB_mod = train(NB_lrn, task1)
NB_task2 = makeClassifTask(data = Test1, target = "Status")
NB_Task.pred = predict(NB_mod, NB_task2, predict.type = "prob")
getConfMatrix(NB_Task.pred)
NB_ROCR_MLR = generateThreshVsPerfData(NB_Task.pred, measures =
list(fpr, tpr, mmce))
plotROCCurves(NB_ROCR_MLR)
NB_ROCRconvert = asROCRPrediction(NB_Task.pred)
NB_ROCR_ROCRcalc = ROCR::performance(NB_ROCRconvert, "tpr", "fpr")
ROCR::plot(NB_ROCR_ROCRcalc, colorize=TRUE,
print.cutoffs.at=seq(0,1,0.1), text.adj=c(-0.2,1.7))
NB_auc = as.numeric(performance(NB_ROCRconvert, "auc")@y.values)
NB_auc
```

The Big Data NB Model Processing

```
... generating 1000 nodes ...
 total number of nodes in initial set                  : 1143
 total number of nodes after removal of identical nodes : 375
 ... computing node means ...
 ... computing node weights ...
 dimension of null space of I                          : 221
 number of selected nodes                              : 31
```

Table 8.11: **The confusion matrix** *of the Big Data NB CF-IPM*

	Predicted as failed firm (0)	**Predicted as existing firm (1)**
Actual failed firm (0)	153	2
Actual existing firm (1)	9	154

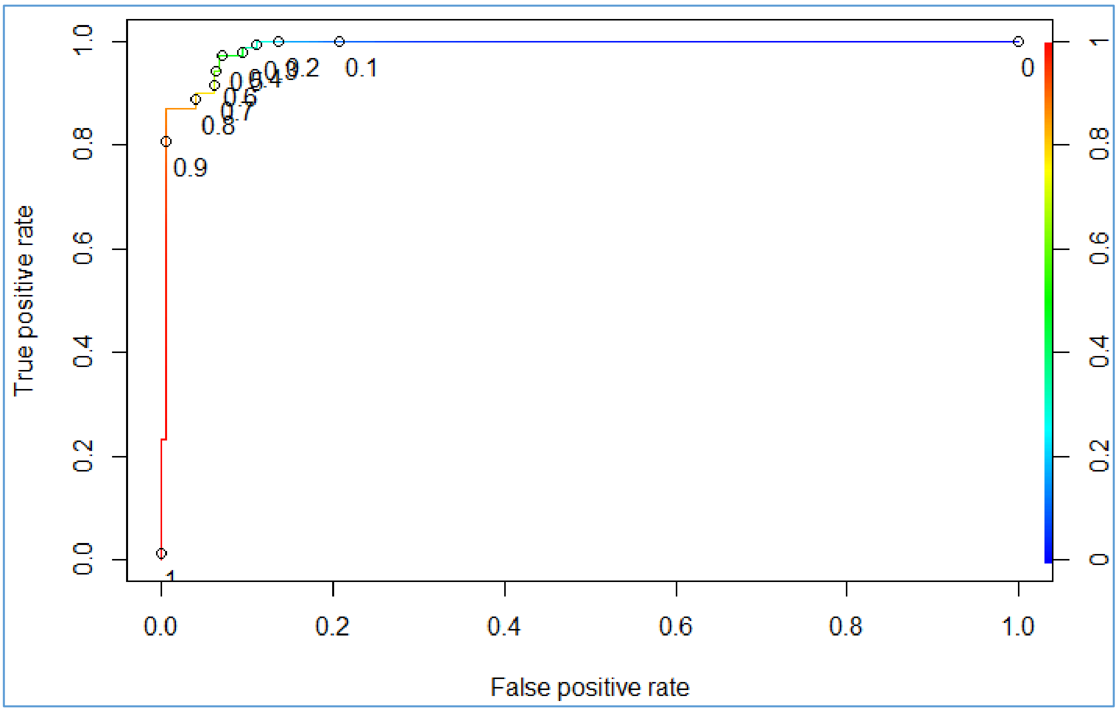


Figure 8.15: **The ROC curve** *of the Big Data NB CF-IPM*

The AUC value of the Big Data NB CF-IPM = 0.9896135

8.5.4.5 The Big Data support vector machine CF-IPM

The code of the Big Data Support Vector Machine (SVM) CF-IPM

```
SVM_lrn = makeLearner("classif.ksvm", predict.type = "prob")
SVM_mod = train(SVM_lrn, task1)
SVM_task2 = makeClassifTask(data = Test1, target = "Status")
SVM_Task.pred = predict(SVM_mod, SVM_task2, predict.type = "prob")
getConfMatrix(SVM_Task.pred)
SVM_ROCR_MLR  =  generateThreshVsPerfData(SVM_Task.pred,  measures  =
list(fpr, tpr, mmce))
plotROCCurves(SVM_ROCR_MLR)
SVM_ROCRconvert = asROCRPrediction(SVM_Task.pred)
SVM_ROCR_ROCRcalc = ROCR::performance(SVM_ROCRconvert, "tpr", "fpr")
ROCR::plot(SVM_ROCR_ROCRcalc,                          colorize=TRUE,
print.cutoffs.at=seq(0,1,0.1), text.adj=c(-0.2,1.7))
SVM_auc = as.numeric(performance(SVM_ROCRconvert, "auc")@y.values)
SVM_auc
```

Table 8.12: **The confusion matrix** *of the Big Data SVM CF-IPM*

	Predicted as failed firm (0)	**Predicted as existing firm (1)**
Actual failed firm (0)	131	24
Actual existing firm (1)	47	116

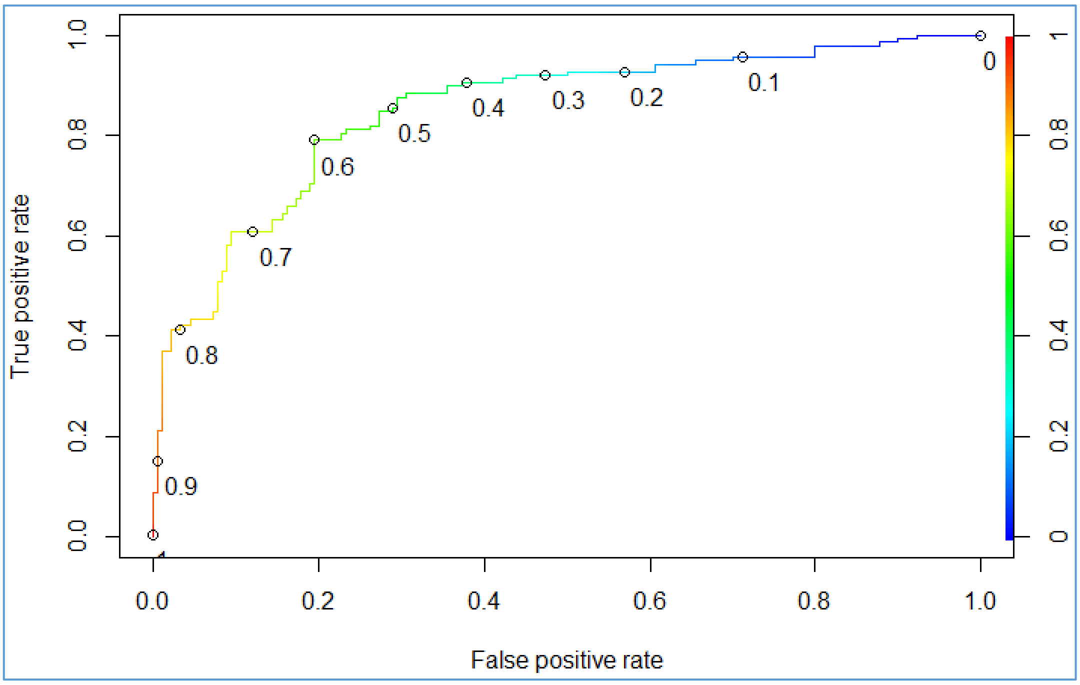


Figure 8.16: **The ROC curve** *of the Big Data SVM CF-IPM*

The AUC value of the Big Data SVM CF-IPM = 0.8498792

8.5.4.6 *The Big data k-nearest neighbour CF-IPM*

<u>The code</u> of the Big Data K-Nearest Neighbour (KNN) CF-IPM

```
KNN_lrn = makeLearner("classif.kknn", predict.type = "prob")
KNN_mod = train(KNN_lrn, task1)
KNN_task2 = makeClassifTask(data = Test1, target = "Status")
KNN_Task.pred = predict(KNN_mod, KNN_task2, predict.type = "prob")
getConfMatrix(KNN_Task.pred)
KNN_ROCR_MLR  =  generateThreshVsPerfData(KNN_Task.pred,  measures  =
list(fpr, tpr, mmce))
plotROCCurves(KNN_ROCR_MLR)
KNN_ROCRconvert = asROCRPrediction(KNN_Task.pred)
KNN_ROCR_ROCRcalc = ROCR::performance(KNN_ROCRconvert, "tpr", "fpr")
ROCR::plot(KNN_ROCR_ROCRcalc,                         colorize=TRUE,
print.cutoffs.at=seq(0,1,0.1), text.adj=c(-0.2,1.7))
KNN_auc = as.numeric(performance(KNN_ROCRconvert, "auc")@y.values)
KNN_auc
```

Table 8.13: **<u>The confusion matrix</u>** *of the Big Data KNN CF-IPM*

	Predicted as failed firm (0)	**Predicted as existing firm (1)**
Actual failed firm (0)	150	5
Actual existing firm (1)	11	152

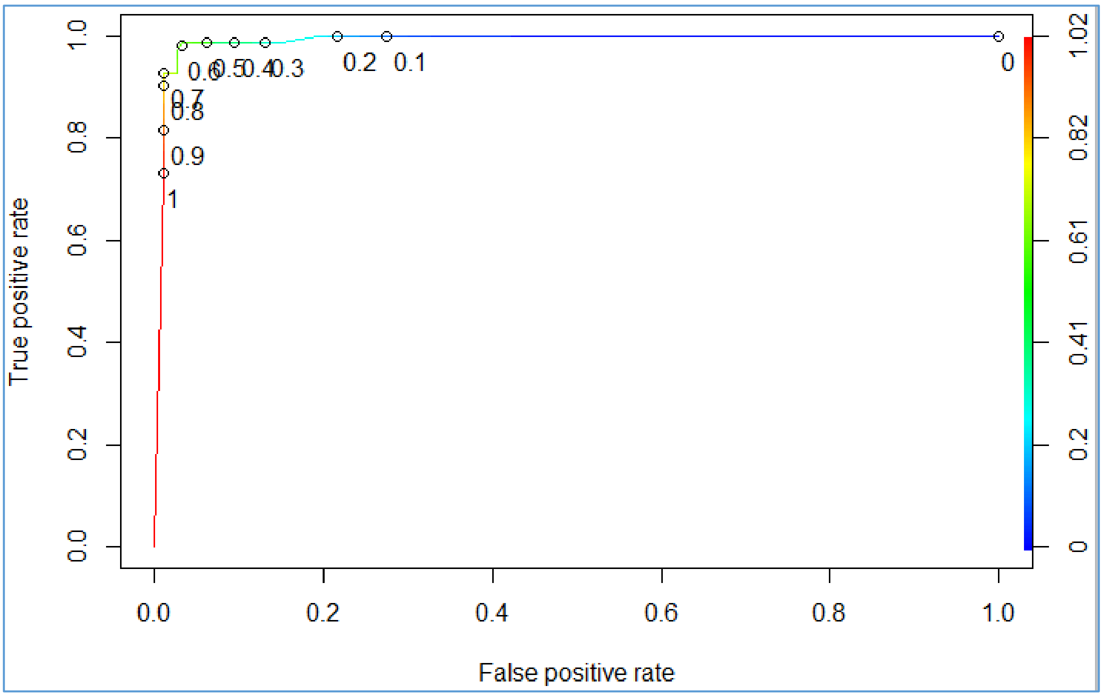


Figure 8.17: **<u>The ROC curve</u>** *of the Big Data KNN CF-IPM*

<u>The AUC value</u> of the Big Data KNN CF-IPM = 0.9877617

8.5.4.7 The Big Data artificial neural network CF-IPM

The code of the Big Data Artificial Neural Network (ANN) CF-IPM

```
NN_lrn = makeLearner("classif.nnet", predict.type = "prob")
NN_mod = train(NN_lrn, task1)
NN_task2 = makeClassifTask(data = Test1, target = "Status")
NN_Task.pred = predict(NN_mod, NN_task2, predict.type = "prob")
getConfMatrix(NN_Task.pred)
NN_ROCR_MLR = generateThreshVsPerfData(NN_Task.pred, measures =
list(fpr, tpr, mmce))
plotROCCurves(NN_ROCR_MLR)
NN_ROCRconvert = asROCRPrediction(NN_Task.pred)
NN_ROCR_ROCRcalc = ROCR::performance(NN_ROCRconvert, "tpr", "fpr")
ROCR::plot(NN_ROCR_ROCRcalc, colorize=TRUE,
print.cutoffs.at=seq(0,1,0.1), text.adj=c(-0.2,1.7))
NN_auc = as.numeric(performance(NN_ROCRconvert, "auc")@y.values)
NN_auc
```

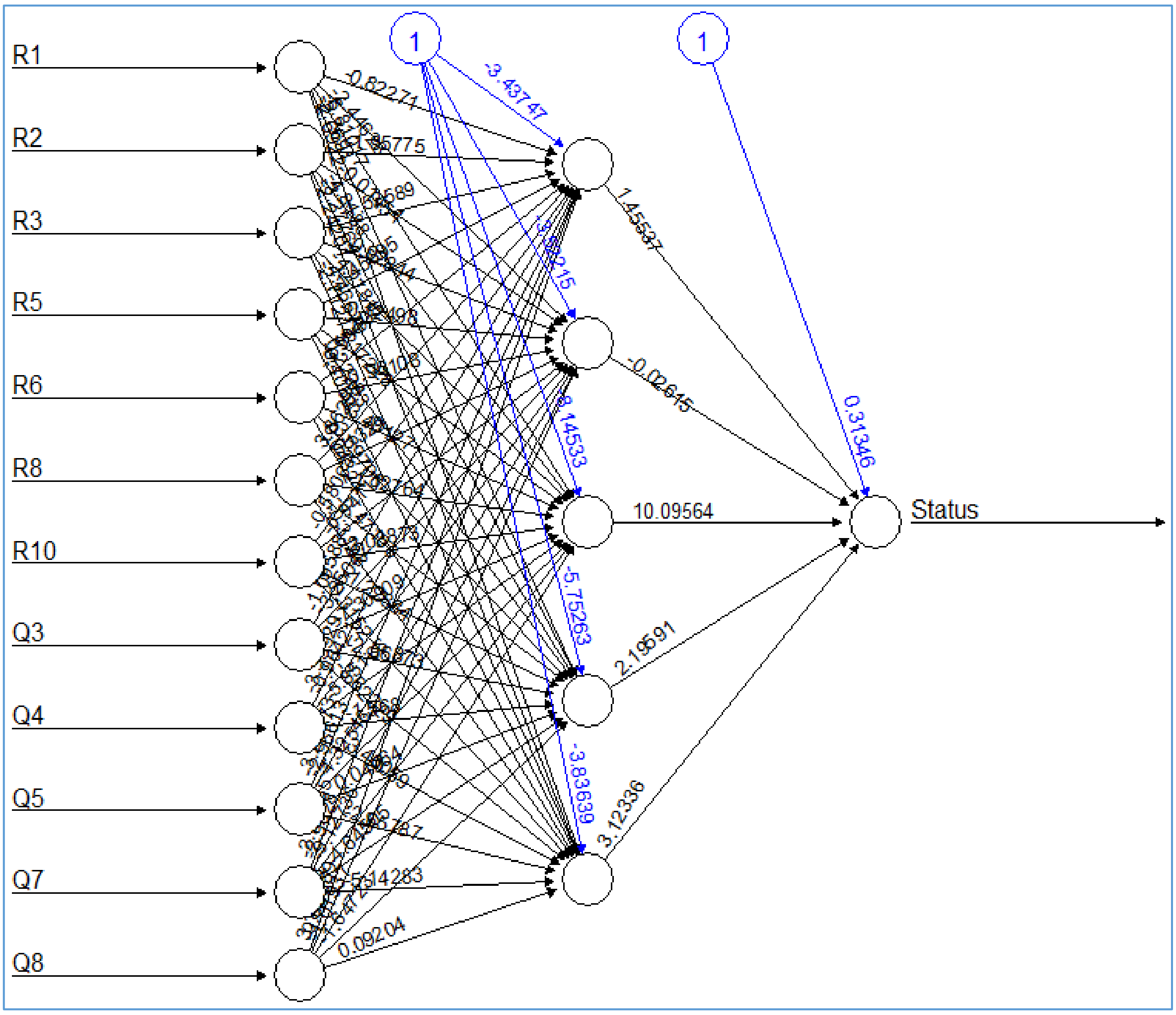


Figure 8.18: **The Big Data ANN model**

The Big Data ANN Model Processing

```
# weights:  34
initial  value 519.513226 
iter  10 value 445.686837
iter  20 value 379.240846
iter  30 value 361.568140
iter  40 value 355.856524
iter  50 value 351.326360
iter  60 value 313.384281
iter  70 value 301.255120
iter  80 value 294.838171
iter  90 value 277.662100
iter 100 value 272.933870
final  value 272.933870 
stopped after 100 iterations
# weights:  34
initial  value 516.019522 
iter  10 value 409.624922
iter  20 value 352.495937
iter  30 value 291.794723
iter  40 value 253.041152
iter  50 value 246.975451
iter  60 value 236.415214
iter  70 value 232.481877
iter  80 value 230.153951
iter  90 value 228.536660
iter 100 value 227.356759
final  value 227.356759 
stopped after 100 iterations
# weights:  34
initial  value 544.862409 
iter  10 value 403.566351
iter  20 value 320.817777
iter  30 value 291.152796
iter  40 value 274.266547
iter  50 value 272.364560
iter  60 value 261.857779
iter  70 value 249.627742
iter  80 value 247.529683
iter  90 value 246.587405
iter 100 value 246.518700
final  value 246.518700 
stopped after 100 iterations
# weights:  34
initial  value 561.494376 
iter  10 value 395.895633
iter  20 value 324.171856
iter  30 value 297.816891
iter  40 value 275.978679
iter  50 value 265.922132
iter  60 value 263.089715
iter  70 value 262.991677
iter  70 value 262.991677
final  value 262.991677 
converged
```

```
# weights:  34
initial  value 525.096494
iter  10 value 417.425941
iter  20 value 322.737804
iter  30 value 295.386522
iter  40 value 264.441029
iter  50 value 249.023055
iter  60 value 244.558910
iter  70 value 244.524399
iter  80 value 244.511842
iter  90 value 244.492043
final  value 244.491997
converged
```

Table 8.14: **The confusion matrix** *of the Big Data ANN CF-IPM*

	Predicted as failed firm (0)	**Predicted as existing firm (1)**
Actual failed firm (0)	133	22
Actual existing firm (1)	19	144

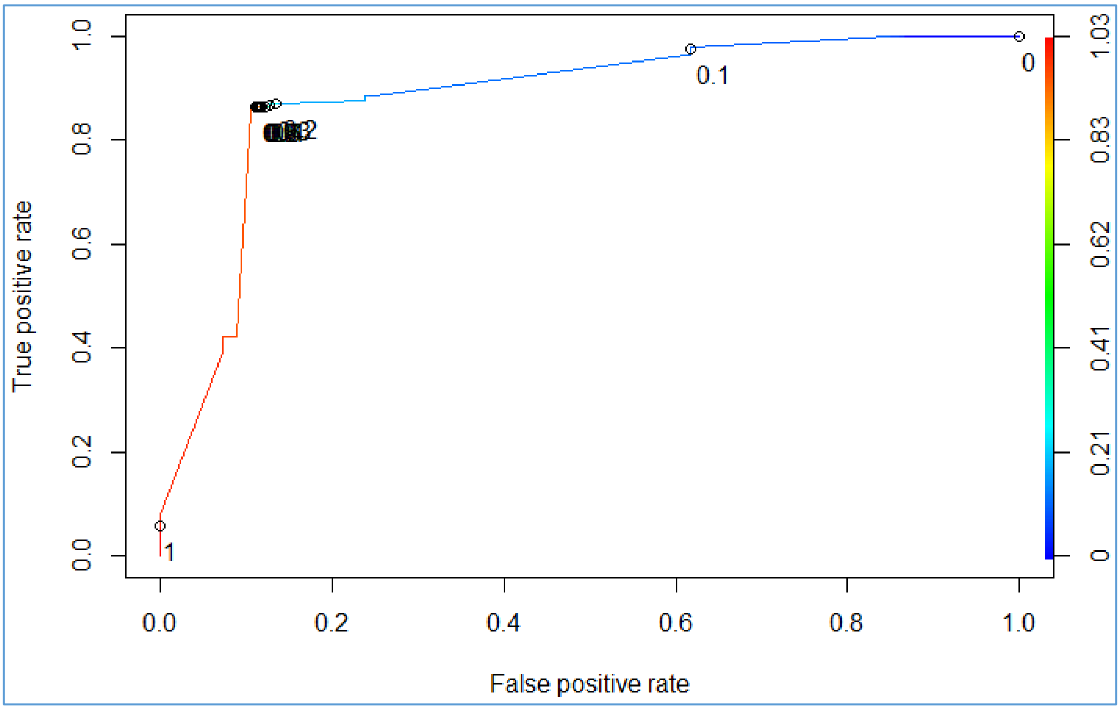


Figure 8.19: **The ROC curve** *of the Big Data ANN CF-IPM*

The AUC value of the Big Data ANN CF-IPM = 0.880394525

8.5.4.8 The Big Data decision tree CF-IPM

The code of the Big Data Decision Tree (DT) CF-IPM

```
DTCART_lrn = makeLearner("classif.rpart", predict.type = "prob")
DTCART_mod = train(DTCART_lrn, task1)
DTCART_task2 = makeClassifTask(data = Test1, target = "Status")
DTCART_Task.pred = predict(DTCART_mod, DTCART_task2, predict.type =
"prob")
getConfMatrix(DTCART_Task.pred)
DTCART_ROCR_MLR    =    generateThreshVsPerfData(DTCART_Task.pred,
measures = list(fpr, tpr, mmce))
plotROCCurves(DTCART_ROCR_MLR)
DTCART_ROCRconvert = asROCRPrediction(DTCART_Task.pred)
DTCART_ROCR_ROCRcalc = ROCR::performance(DTCART_ROCRconvert, "tpr",
"fpr")
ROCR::plot(DTCART_ROCR_ROCRcalc,                   colorize=TRUE,
print.cutoffs.at=seq(0,1,0.1), text.adj=c(-0.2,1.7))
DTCART_auc      =      as.numeric(performance(DTCART_ROCRconvert,
"auc")@y.values)
DTCART_auc
```

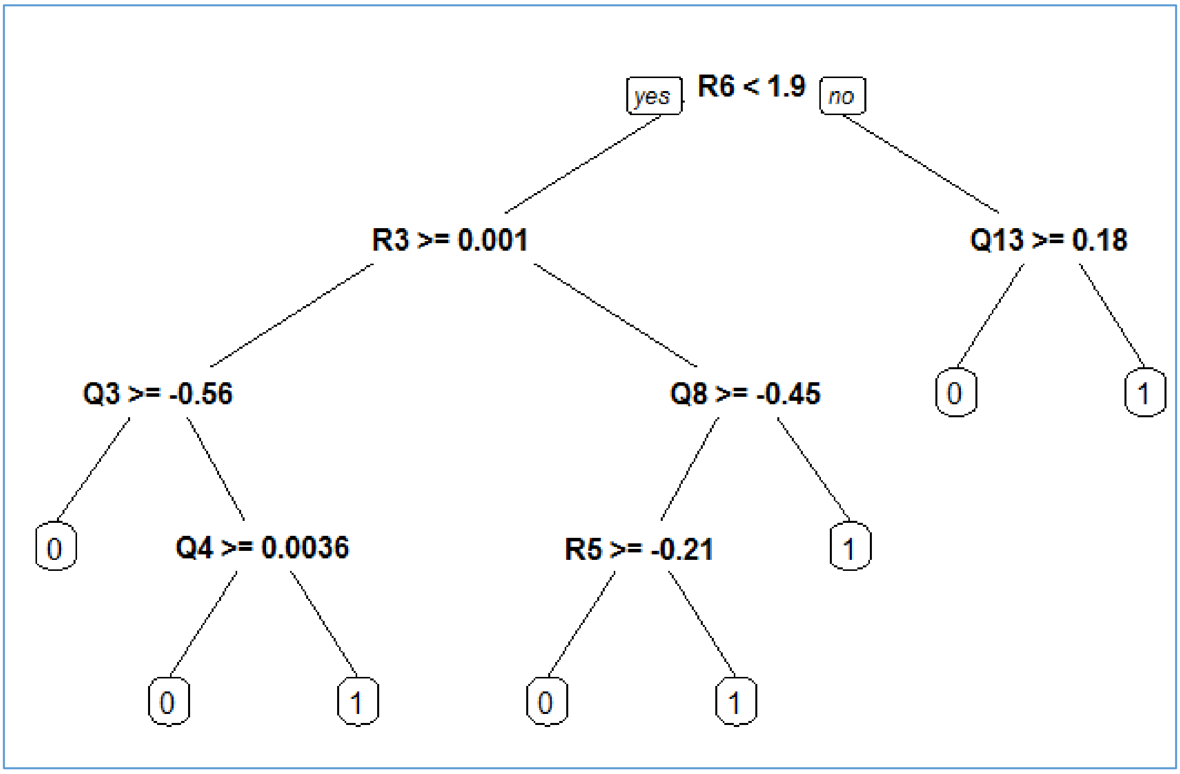


Figure 8.20: **The Big Data DT model**

Table 8.15: **The confusion matrix** *of the Big Data DT CF-IPM*

	Predicted as failed firm (0)	**Predicted as existing firm (1)**
Actual failed firm (0)	150	15
Actual existing firm (1)	19	144

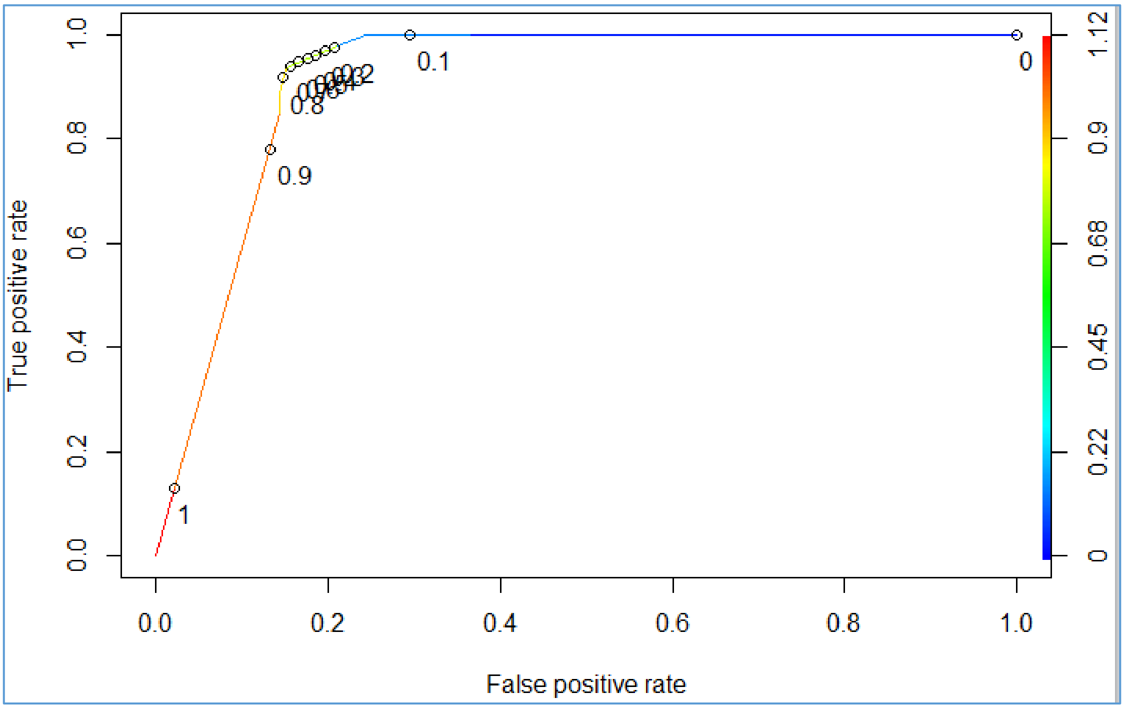


Figure 8.21: **The ROC curve** *of the Big Data DT CF-IPM*

The AUC value of the Big Data DT CF-IPM = 0.9432648953

8.5.4.9 The Big Data random forest CF-IPM

The code of the Big Data Random Forest (RF) CF-IPM

```
RF_lrn = makeLearner("classif.randomForest", predict.type = "prob")
RF_mod = train(RF_lrn, task1)
RF_task2 = makeClassifTask(data = Test1, target = "Status")
RF_Task.pred = predict(RF_mod, RF_task2, predict.type = "prob")
getConfMatrix(RF_Task.pred)
RF_ROCR_MLR = generateThreshVsPerfData(RF_Task.pred, measures =
list(fpr, tpr, mmce))
plotROCCurves(RF_ROCR_MLR)
RF_ROCRconvert = asROCRPrediction(RF_Task.pred)
RF_ROCR_ROCRcalc = ROCR::performance(RF_ROCRconvert, "tpr", "fpr")
ROCR::plot(RF_ROCR_ROCRcalc, colorize=TRUE,
print.cutoffs.at=seq(0,1,0.1), text.adj=c(-0.2,1.7))
RF_auc = as.numeric(performance(RF_ROCRconvert, "auc")@y.values)
RF_auc
```

Table 8.16: **The confusion matrix** *of the Big Data RF CF-IPM*

	Predicted as failed firm (0)	**Predicted as existing firm (1)**
Actual failed firm (0)	155	0
Actual existing firm (1)	3	160

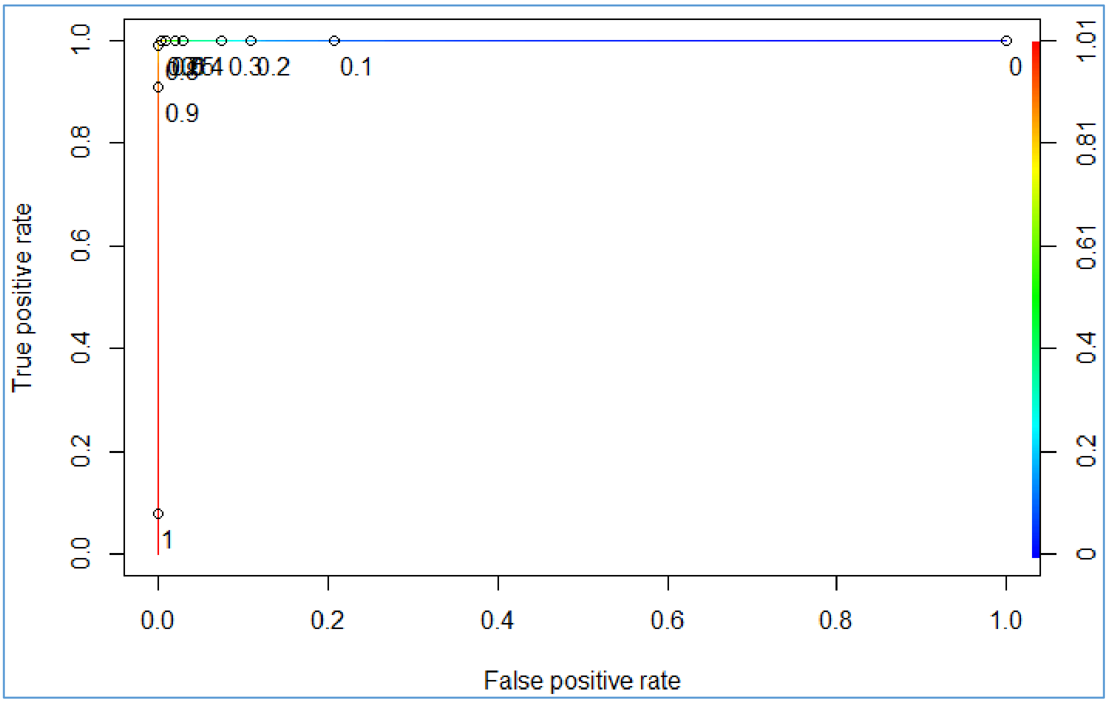


Figure 8.22: **The ROC curve** *of the Big Data RF CF-IPM*

The AUC value of the Big Data RF CF-IPM = 1.0

8.5.4.10 The Big Data bart machine CF-IPM

<u>The code</u> of the Big Data Bart Machine (BM) CF-IPM

```
BM_lrn = makeLearner("classif.bartMachine", predict.type = "prob")
BM_mod = train(BM_lrn, task1)
BM_task2 = makeClassifTask(data = Test1, target = "Status")
BM_Task.pred = predict(BM_mod, BM_task2, predict.type = "prob")
getConfMatrix(BM_Task.pred)
BM_ROCR_MLR = generateThreshVsPerfData(BM_Task.pred, measures = 
list(fpr, tpr, mmce))
plotROCCurves(BM_ROCR_MLR)
BM_ROCRconvert = asROCRPrediction(BM_Task.pred)
BM_ROCR_ROCRcalc = ROCR::performance(BM_ROCRconvert, "tpr", "fpr")
ROCR::plot(BM_ROCR_ROCRcalc, colorize=TRUE, 
print.cutoffs.at=seq(0,1,0.1), text.adj=c(-0.2,1.7))
BM_auc = as.numeric(performance(BM_ROCRconvert, "auc")@y.values)
BM_auc
```

<u>The Big Data BM Model Processing</u>

```
bartMachine vars checked...
bartMachine java init...
bartMachine factors created...
bartMachine before preprocess...
bartMachine after preprocess... 10 total features...
bartMachine sigsq estimated...
bartMachine training data finalized...
Now building bartMachine for classification ...Missing data 
feature ON.
building BART with mem-cache speedup...
Iteration 100/1250  mem: 26.8/519MB
Iteration 200/1250  mem: 36.5/519MB
Iteration 300/1250  mem: 27.6/519MB
Iteration 400/1250  mem: 37.8/519MB
Iteration 500/1250  mem: 34.2/519MB
Iteration 600/1250  mem: 43.4/519MB
Iteration 700/1250  mem: 50/519MB
Iteration 800/1250  mem: 54.7/519MB
Iteration 900/1250  mem: 46.5/519MB
Iteration 1000/1250  mem: 61.8/519MB
Iteration 1100/1250  mem: 56.9/519MB
Iteration 1200/1250  mem: 72.3/519MB
done building BART in 6.228 sec


burning and aggregating chains from all threads... done
evaluating in sample data...done
```

Table 8.17: **<u>The confusion matrix</u>** *of the Big Data BM CF-IPM*

	Predicted as failed firm (0)	**Predicted as existing firm (1)**
Actual failed firm (0)	154	1
Actual existing firm (1)	5	158

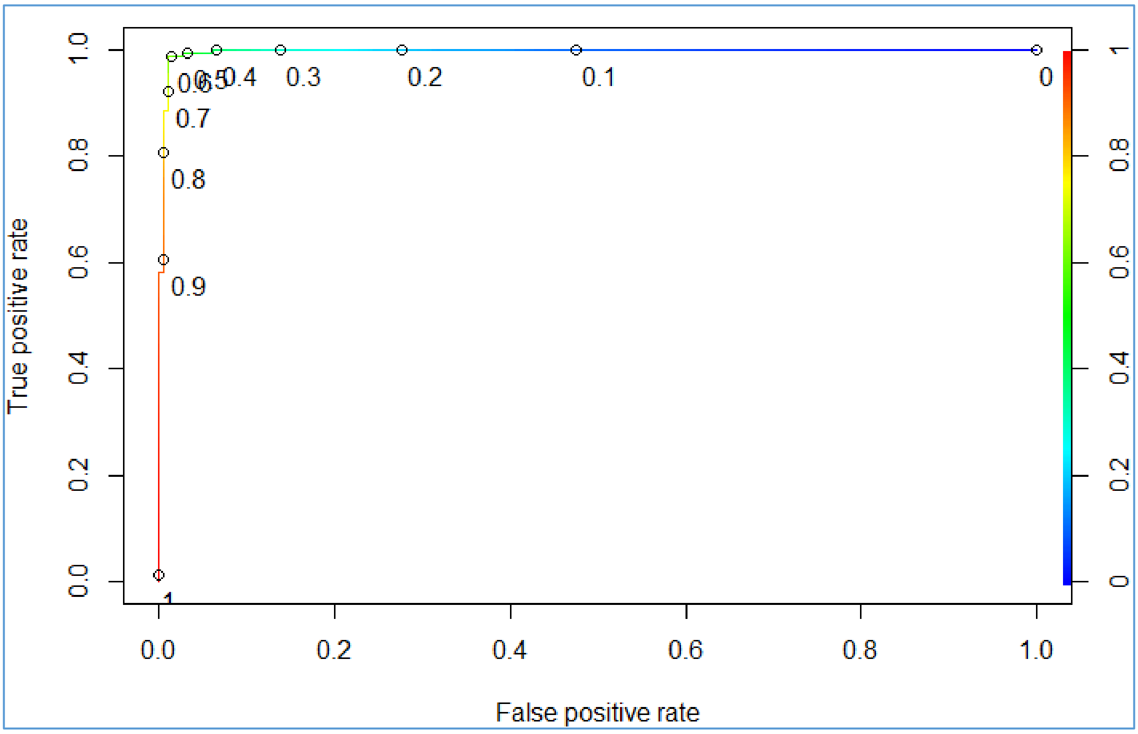


*Figure 8.23:***<u>The ROC curve</u>** *of the Big Data BM CF-IPM*

<u>The AUC value</u> of the Big Data BM CF-IPM = 0.9965378422

8.5.4.11 The Big Data adaptive boosting CF-IPM

The code of the Big Data Adaptive Boosting (AB) CF-IPM

```
AB_lrn = makeLearner("classif.ada", predict.type = "prob")
AB_mod = train(AB_lrn, task)
AB_task2 = makeClassifTask(data = Test, target = "Status")
AB_Task.pred = predict(AB_mod, AB_task2, predict.type = "prob")
getConfMatrix(AB_Task.pred)
AB_ROCR_MLR = generateThreshVsPerfData(AB_Task.pred, measures =
list(fpr, tpr, mmce))
plotROCCurves(AB_ROCR_MLR)
AB_ROCRconvert = asROCRPrediction(AB_Task.pred)
AB_ROCR_ROCRcalc = ROCR::performance(AB_ROCRconvert, "tpr", "fpr")
ROCR::plot(AB_ROCR_ROCRcalc, colorize=TRUE,
print.cutoffs.at=seq(0,1,0.1), text.adj=c(-0.2,1.7))
AB_auc = as.numeric(performance(AB_ROCRconvert, "auc")@y.values)
AB_auc
```

Table 8.18: **The confusion matrix** *of the Big Data AB CF-IPM*

	Predicted as failed firm (0)	**Predicted as existing firm (1)**
Actual failed firm (0)	154	1
Actual existing firm (1)	1	162

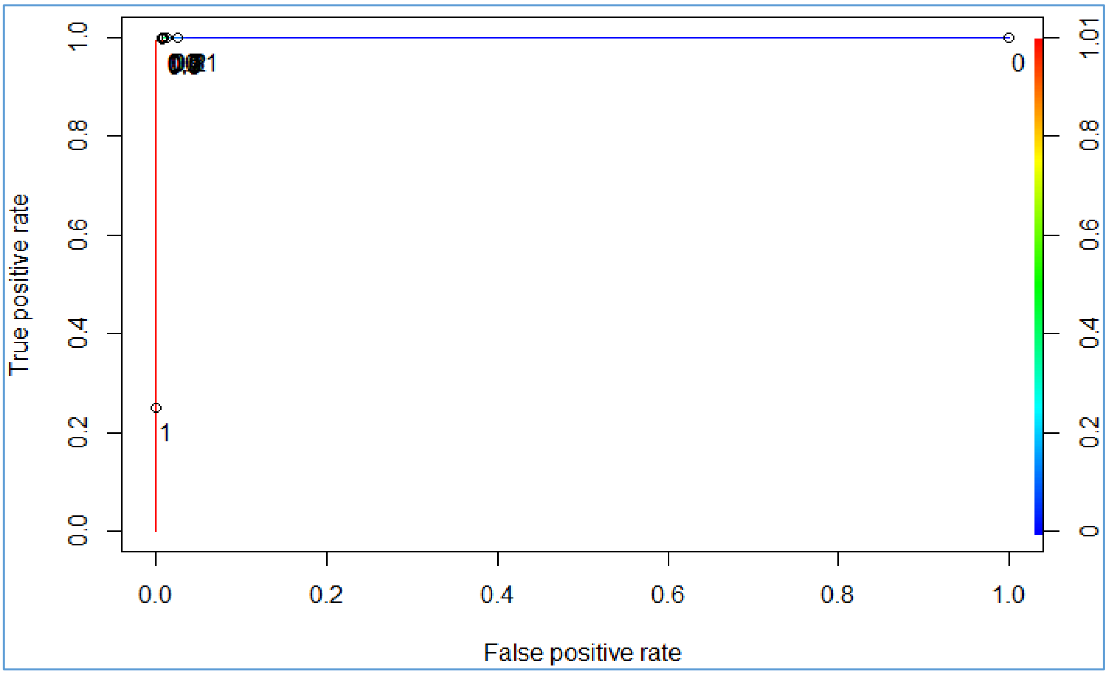


Figure 8.24: **The ROC curve** *of the Big Data AB CF-IPM*

The AUC value of the Big Data AB CF-IPM = 0.9999604196

8.5.4.12 The Big Data propositional rule learner CF-IPM

The code of the Big Data Propositional Rule Learner (PRL) CF-IPM

```
PRL_lrn = makeLearner("classif.JRip", predict.type = "prob")
PRL_mod = train(PRL_lrn, task1)
PRL_task2 = makeClassifTask(data = Test1, target = "Status")
PRL_Task.pred = predict(PRL_mod, PRL_task2, predict.type = "prob")
getConfMatrix(PRL_Task.pred)
PRL_ROCR_MLR = generateThreshVsPerfData(PRL_Task.pred, measures =
list(fpr, tpr, mmce))
plotROCCurves(PRL_ROCR_MLR)
PRL_ROCRconvert = asROCRPrediction(PRL_Task.pred)
PRL_ROCR_ROCRcalc = ROCR::performance(PRL_ROCRconvert, "tpr",
"fpr")
ROCR::plot(PRL_ROCR_ROCRcalc, colorize=TRUE,
print.cutoffs.at=seq(0,1,0.1), text.adj=c(-0.2,1.7))
PRL_auc = as.numeric(performance(PRL_ROCRconvert, "auc")@y.values)
PRL_auc
```

Table 8.19: **The confusion matrix** *of the Big Data PRL CF-IPM*

	Predicted as failed firm (0)	**Predicted as existing firm (1)**
Actual failed firm (0)	149	6
Actual existing firm (1)	13	150

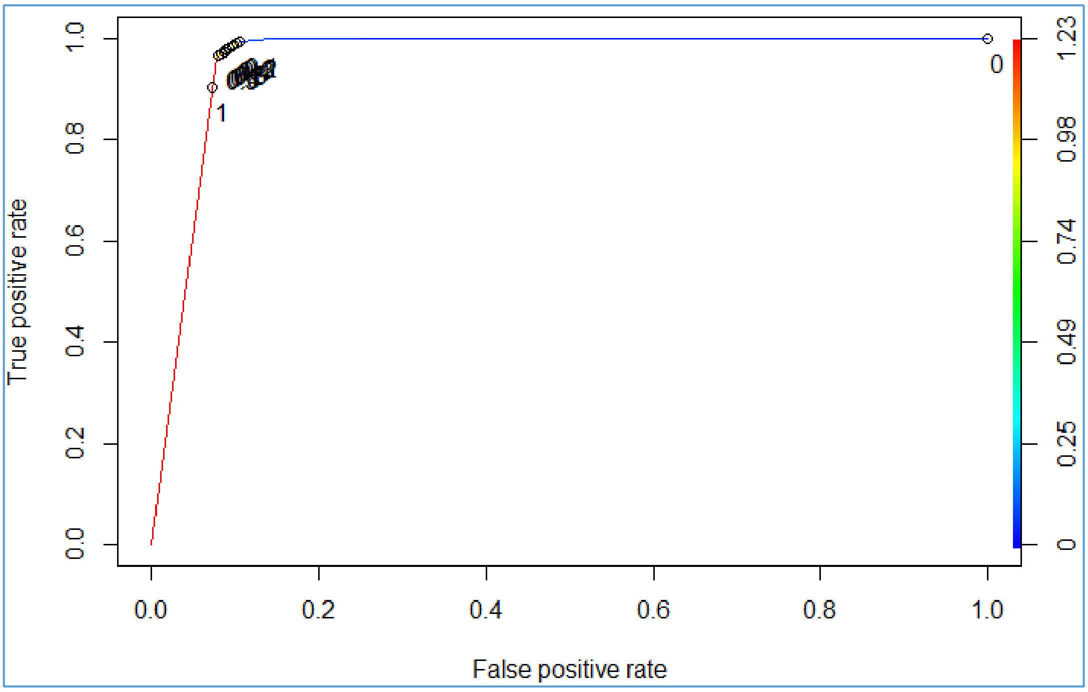


Figure 8.25: **The ROC curve** *of the Big Data PRL CF-IPM*

The AUC value of the Big Data PRL CF-IPM = 0.9589573269

8.5.4.13 The Big Data kohonen CF-IPM

<u>The code</u> of the Big Data Kohonen (KHN) CF-IPM

```
KHN_lrn = makeLearner("classif.bdk", predict.type = "prob")
KHN_mod = train(KHN_lrn, task1)
KHN_task2 = makeClassifTask(data = Test1, target = "Status")
KHN_Task.pred = predict(KHN_mod, KHN_task2, predict.type = "prob")
getConfMatrix(KHN_Task.pred)
KHN_ROCR_MLR = generateThreshVsPerfData(KHN_Task.pred, measures =
list(fpr, tpr, mmce))
plotROCCurves(KHN_ROCR_MLR)
KHN_ROCRconvert = asROCRPrediction(KHN_Task.pred)
KHN_ROCR_ROCRcalc = ROCR::performance(KHN_ROCRconvert, "tpr", "f")
ROCR::plot(KHN_ROCR_ROCRcalc, colorize=TRUE,
print.cutoffs.at=seq(0,1,0.1), text.adj=c(-0.2,1.7))
KHN_auc = as.numeric(performance(KHN_ROCRconvert, "auc")@y.values)
KHN_auc
```

Table 8.20: **<u>The confusion matrix</u>** *of the Big Data KHN CF-IPM*

	Predicted as failed firm (0)	**Predicted as existing firm (1)**
Actual failed firm (0)	135	20
Actual existing firm (1)	18	145

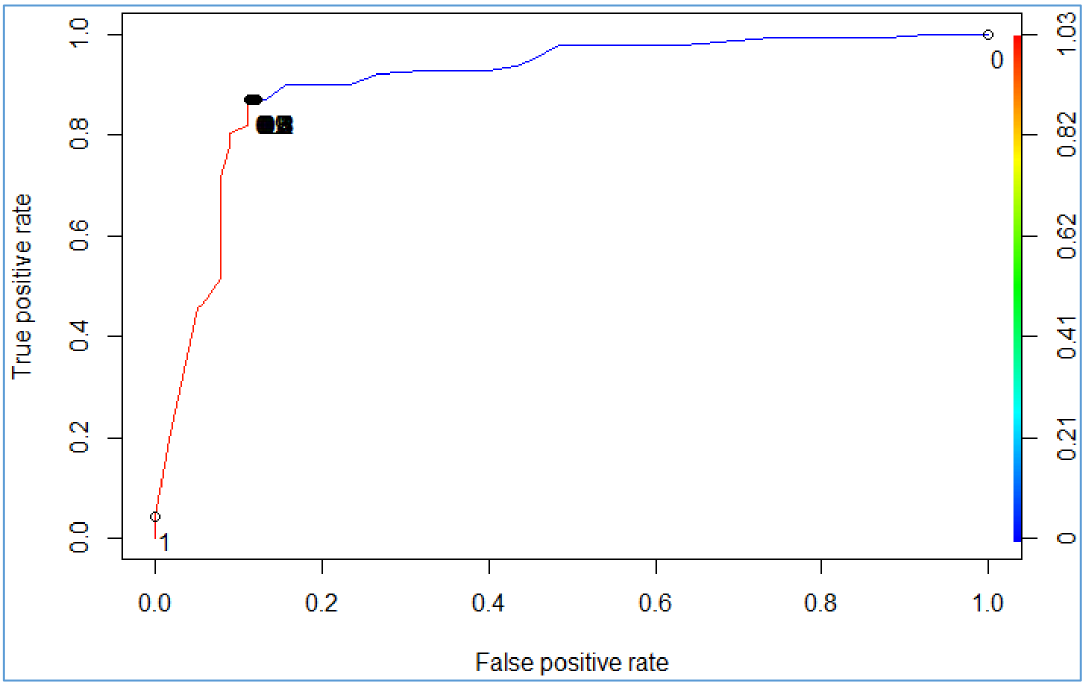


Figure 8.26: **<u>The ROC curve</u>** *of the Big Data KHN CF-IPM*

<u>The AUC value</u> of the Big Data KHN CF-IPM = 0.9046900161

8.5.5 Summary of results of CF-IPMs developed

Table 8.21 is a presentation of the comparison of the 13 Big Data CF-IPMs developed. These models (i.e. Big Data CF-IPMS) will be discussed in the next chapter. As explained earlier in subsection 8.5.3, the calculations for overall accuracy, specificity, sensitivity and, Types I and II error are based on the confusion matrix of each model.

Table 8.21: Summary of the results of the Big Data CF-IPMs developed

S/N	**Tool/Algorithm**	**True negatives**	**True positives**	**AUC**	**Overall accuracy**	**Sensitivity**	**Specificity**	**Type I error**	**Type II error**
1)	**Big Data Linear Discriminant Analysis**	116	119	0.82109	0.739	0.730	0.748	0.252	0.270
2)	**Big Data Quadratic Discriminant Analysis**	110	117	0.82105	0.714	0.718	0.710	0.290	0.282
3)	**Big Data Logistic Regression**	110	117	0.82105	0.714	0.718	0.710	0.290	0.282
4)	**Big Data Naïve Bayes**	153	154	0.98961	0.965	0.945	0.987	0.013	0.055
5)	**Big Data Support Vector Machine**	131	116	0.84987	0.777	0.712	0.845	0.155	0.288
6)	**Big Data K-Nearest Neighbour**	150	152	0.98776	0.950	0.933	0.968	0.032	0.067
7)	**Big Data Artificial Neural Network**	133	144	0.88039	0.871	0.883	0.858	0.142	0.117
8)	**Big Data Decision Tree**	145	139	0.91326	0.925	0.883	0.968	0.032	0.117
9)	**Big Data Random Forest**	155	160	1.00000	0.991	0.982	1.000	0.000	0.018
10)	**Big Data Bart Machine**	154	158	0.99653	0.981	0.969	0.994	0.006	0.031
11)	**Big Data Adaptive Boosting**	154	162	0.99996	0.994	0.994	0.994	0.006	0.006
12)	**Big Data Propositional Rule Learner**	149	166	0.95895	0.940	0.922	0.964	0.036	0.078
13)	**Big Data Kohonen**	135	160	0.90469	0.881	0.889	0.870	0.130	0.111

8.6 Chapter summary

The development of the Big Data Analytics CF-IPMs is detailed in this chapter. The ‘R’ programming language was used to develop the CF-IPMs. The AWS EC2 and six ‘Spot Instances’ were used to set up the Big Data Analytics platform with Apache Spark computation engine. Five of the Instances were run as Hadoop DataNodes and one as Hadoop NameNode which was subsequently configured as Spark Master. The data was uploaded to the DataNodes and split into model (i.e. CF-IPM) training (70%) and testing (30%) data. The variables to be used to develop the CF-IPMs were selected using a voting system on the results of 11 variables selection techniques including information gain, Kruskal test, minimum redundancy, mRMR, chi-squared, gain ratio, among others.

The selected variables were used to develop 13 different Big Data Analytic CF-IPMs using Apache Spark with 13 predictive tools/algorithms including ANN, SVM, KNN, RF, AB, KNN, LR, LDA, QDA, NB DT, PRL and KHN. The details presented with each model include the execution code, the model (where physical model exist), the model processing (where ‘R’ displays it), confusion matrix, ROC curve, and AUC value. Finally, a summary table displaying the true negatives, true positives, AUC, overall accuracy, sensitivity, specificity, Type I error and Type II error, for each of the 13 models is given.

A discussion of the results was presented in chapter nine. The results discussed include analytical results, the variables proven to affect insolvency of construction firms, and implication of the result on theory.

CHAPTER NINE

9.0 MODEL SELECTION (RESULT) AND DISCUSSION

9.1 Chapter introduction

This chapter is a presentation of the results of the research alongside its discussion. The results discussed include those of the 13 developed Big Data construction firms insolvency prediction models (CF-IPM) and the factors produced by the best Big Data CF-IPM. Section 9.2 is an explanation of how the best CF-IPM was selected based on three selection criteria that are important to the aim of the research: (i) accuracy, (ii) error type levels and (iii) transparency. Subsection 9.2.1 is used to explain how the best CF-IPMs based on accuracy were selected. Subsections 9.2.2 and 9.2.3 are used to explain how the best CF-IPMs based on error type levels and transparency were selected respectively. Subsection 9.2.4 is a presentation of the best CF-IPM which is the Decision Tree CF-IPM. Section 9.3 is a description of the most important factors affecting (in)solvency of construction firms based on the factors (i.e. variables) produced by the best CF-IPM. Subsections 9.3.1 and 9.3.2 were used to describe the quantitative and qualitative factors respectively. Section 9.4 is an explanation of the implication of the research on theory in terms of its support for a multi-theory basis for the (in)solvency of construction firms. Section 9.5 is a summary of the chapter.

9.2 Selection of the best big data CF-IPM

Selecting the best out of a number of models is absolutely dependent on the intention of the developer which is based on the aim of the user. A financier, for example, will be mainly interested in the accuracy of a CF-IPM to decide whether or not to give a loan to the construction firm. In the research, the main target users were construction firm’s owners and the overall intention is to reduce the rate of failure of firms in the construction industry. Accuracy, transparency and a reduction in Type I error are thus vital to the selection of the best CF-IPM.

The Big Data CF-IPM of choice in the research is selected based on the afore-mentioned three attributes. The summary of the results of the Big Data CF-IPMs developed are re-presented here in Table 9.1 with the overall accuracy, sensitivity, specificity and error types expressed as percentages. Table 9.1 also gives the transparency condition of each of the Big Data CF-IPMS

Table 9.1: Summary of the results of the Big Data CF-IPMs developed and their transparency condition

	Tool/Algorithm	**True -tives**	**True +tives**	**AUC**	**Overall accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**	**Type I error (%)**	**Type II error (%)**	**Transparent**
1.	**Big Data Linear Discriminant Analysis (LDA)**	116	119	0.82109	73.90	73.01	74.84	25.16	26.99	Yes
2.	**Big Data Quadratic Discriminant Analysis (QDA)**	110	117	0.82105	71.38	71.78	70.97	29.03	28.22	Yes
3.	**Big Data Logistic Regression (LR)**	110	117	0.82105	71.38	71.78	70.97	29.03	28.22	Yes
4.	**Big Data Naïve Bayes (NB)**	153	154	0.98961	96.54	94.48	98.71	1.29	5.52	No
5.	**Big Data Support Vector Machine (SVM)**	131	116	0.84987	77.67	71.17	84.52	15.48	28.83	No
6.	**Big Data K-Nearest Neighbour (KNN)**	150	152	0.98776	94.97	93.25	96.77	3.23	6.75	No
7.	**Big Data Artificial Neural Network (ANN)**	133	144	0.88039	87.11	88.34	85.81	14.19	11.66	No
8.	**Big Data Decision Tree (DT)**	145	139	0.94326	92.45	88.34	96.77	3.23	11.66	Yes
9.	**Big Data Random Forest (RF)**	155	160	1.00000	99.06	98.16	100.0	0.00	1.84	No
10.	**Big Data Bart Machine BM)**	154	158	0.99653	98.11	96.93	99.35	0.65	3.07	No
11.	**Big Data Adaptive Boosting (AB)**	154	162	0.99996	99.37	99.39	99.35	0.65	0.61	No
12.	**Big Data Propositional Rule Learner (PRL)**	149	166	0.95895	94.03	92.02	96.13	3.87	7.98	No
13.	**Big Data Kohonen (KHN)**	135	160	0.90469	88.05	88.96	87.10	12.90	11.04	No

No.: Number *-tives: negatives* *+tives: positives*

9.2.1 Model accuracy

As with any target user, accuracy is vital for construction firm owners (or top management team) since too many wrong predictions make a CF-IPM unreliable and more or less useless for construction firms (i.e. the target users). For a test data set that contains absolutely equal number of failed and existing construction firms, a simple rough prediction stating that all the firms exist (or have failed) will give a 50% accuracy; an accuracy level unacceptable by the standards of any set of target users. In fact, it is a gamble level accuracy. A CF-IPM thus has to do much better with accuracy levels of well over 85%, or even 90%, if it has to influence key business decisions. The cost of misclassifying a single construction firm alone can be devastating. A construction firm wrongly classified as failing might end up truly failing simply because clients can avoid awarding contracts to such firm on the premise of the misclassification.

Overall, more than 50% of the 13 CF-IPMs developed have an accuracy of over 90% on test data, depicting a very good model development process. Table 9.1 shows that CF-IPMs developed with ensemble AI tools (RF, AB and BM) are in particular extremely accurate with accuracy values of over 98%. RF CF-IPM and AB CF-IPM are more accurate than BM CF-IPM as they have over 99% accuracy. Although AB CF-IPM seems to be 0.3 percent more accurate the RF CF-IPM, the RF CF-IPM is the best model in terms of accuracy performance because its AUC value is 1.0, depicting a perfect model. The AUC value of 1.0 implies that the RF CF-IPM is totally trusted to give a superb performance on any new data beyond the test data. The AB CF-IPM is however not a distant second best in terms of accuracy as its AUC value of 0.99996 is as close to 1.0 as it can get. All other artificial intelligence (AI) tools but SVM, ANN and KHN produced CF-IPMs with high accuracy (i.e. over 90% overall accuracy). Overall any of the seven CF-IPMs with high accuracy (see Table 9.1) can be selected in the research depending on how they satisfy the two remaining selection criteria. Although the Type I error of the all the highly accurate CF-IPMs (apart from RF CF-IPM) could be reduced by moving the threshold. This act will increase an already ‘not too satisfactory’ Type II error of some of the CF-IPMs (e.g. ANN, KHN, among others) hence there is no point doing this.

The statistical tools’ CF-IPMs (i.e. LDA, QDA and LR) appear to be far behind the AI tools’ CF-IPMs apart from SVM which, alongside ANN, which inexplicably underperformed as they are widely known to be very accurate AI tools (Liang, Tsai, and Wu, 2015; Tseng and Hu, 2010; Yeh, Chi, and Lin, 2014, among others). These tools’ (i.e. SVM and ANN) underperformance could probably be attributed to two elements of the CF-IPM development process in the research. One element is the tools' possible inability to handle qualitative variables since all studies that proved their high accuracy used only quantitative variables. Such possibility will mean a reduction in the fitness of these tools to developing CF-IPMs. This is proclaimed because it has long been established that the fitness of any insolvency prediction model (IPM) to construction firms is dependent on its use of qualitative variables among other factors (Arditi, Koksal, and Kale, 2000; Hall, 1994; Horta and Camanho, 2013; Kale and Arditi, 1999; Kangari, 1988, among others). Another element is that SVM and ANN look like they underperformed because they were compared some stronger AI tools like RF and AB that use ensemble methods (see subsection 8.5.2). This is, however, less likely to be the case since other standard (i.e. non-ensemble) AI tools like KNN and NB returned results with a very high accuracy of over 90%. More so, the poor overall accuracy of SVM in particular barely distinguished it from statistical tools. Although KHN did not perform much better than ANN, it cannot be singled out for criticism like ANN and SVM since it is not popularly known for high accuracy in developing IPMs. In fact, it is not popular with IPM developers.

9.2.2 Model error type levels

The cost of error is another CF-IPM feature of great interest to construction firms and other potential users. Type I error, which is the costlier, happens when a failing firm is wrongly predicted as being healthy. The main cost of this error stems from the fact that it deceives a failing construction firm into thinking it is healthy thereby causing the top management team (TMT) to carry out operations as normal without seeking redress to the firm's situation. This error type will thus not aid the reduction in the number of failing construction firms, as advocated by the research. In fact, it will increase the number in that, a construction firm's TMT that senses the firm is in some troubles will be wrongly reassured that there is no problem after a Type I error misclassification. Type II error, on the hand, will give a healthy firm's management a wrong feeling of impending failure thereby causing it to take remedial steps. Such steps can make the construction firm even stronger. Type II error, however, has its own cost though lesser than Type I. For example, although it is not a common practice, a construction firm's management can easily decide to shut down in order to minimise loss if the firm is predicted to be failing (Kuo, 2013). Type II error can thus cause an existing construction firm to fail, though this happens very sparingly.

The CF-IPMs from eight tools performed well in terms of reducing Type I error. RF CF-IPM leads the way with zero Type I error. Although AB's Type I error is greater than its Type II, both errors are negligible at values less than 0.7%. In fact, the AB CF-IPM misclassified (or wrongly predicted) only one existing and failed firm and produced the joint second lowest Type I error along with BM CF-IPM. The CF-IPMS developed with NB, KNN, DT, and PRL all produced Type I error values of less than 4%, surpassing the 10% acceptable error benchmark used in the proposed solution, making them very strong models.

The statistical tools (i.e. LDA, QDA and LR) alongside SVM, ANN and KHN appear to be the poorest regarding reducing Type I error. CF-IPMs developed with each of these tools produced Type I error values of above 10%, depicting less than 90% prediction accuracy on failed construction firms. The statistical tools are in particular worse off with Type I error values of above 25% or even close to 30%. Of the poor tools mentioned, only LDA and SVM produced lower Type I error than Type II.

9.2.3 Model transparency

For construction firm owners, transparency is the most important feature of a CF-IPM after accuracy, especially for construction firms predicted to be failing. A non-transparent CF-IPM with 100% accuracy will not do a construction firm predicted to be failing many favours after the prediction. Although the firm TMT will understand the imminent danger, they would not have been helped with identifying the problems. The firm will eventually set into panic mode, and making decisions will become extremely difficult. A transparent model will do a lot better by displaying the exact factors (or variables) causing problems for the construction firm, making the task of preventing the impending failure somewhat easier.

Transparency of a model normally comes in one of two ways; either as an equation or as an interpretable network diagram. Of the 13 tools used to build the 13 CF-IPMs, only four are transparent enough to understand their results. These four include the three statistical tools and the DT CF-IPM. While the statistical tools CF-IPM produced equations, the DT CF-IPM produced an interpretable network diagram (see Figure 8.20). The fact that the QDA and LR produced exactly the same model (see sub-subsections 8.5.4.2 and 8.5.4.3) is not too surprising since quadratic equation is also referred to as logistic difference equation. The remaining nine other tools CF-IPMS are not transparent enough to allow understanding of result. Although ANN produced a network diagram, the diagram is not interpretable. This is why ANN, as well as SVM, are known as ‘black box’ tools. Although the ensemble classifier models (i.e. RF and AB CF-IPMs) are not transparent, their likely set of most important variables can be gotten from the CForest variable selection technique. This is because CForest is also an ensemble technique and the three (i.e. CForest, RF and AB) are most likely to select the same set of variables as the most important, given a group of variables.

The coefficients assigned to variables in the equation produced by the statistical methods represent the importance of the variables. The higher the coefficient value of a variable, the more important that variable is. The equations of the statistical tools’ CF-IPMs, as presented in sub-subsections 8.5.4.1, 8.5.4.2 and 8.5.4.3show that the models found only quantitative variables to be important. While the coefficient assigned to all qualitative variables by the model were below 1.0 (negligible), the ones assigned to the best quantitative variables in the model ranged between 28 and 131. This proves the statistical models are unfit to produce very effective IPMs for construction businesses since they cannot handle qualitative variables which are very important for predicting insolvency of construction firms as highlighted earlier. This is probably why their results are exceptionally poor relatively.

9.2.4 Decision tree CF-IPM as the Choice Model in the research

The choice CF-IPM of the research is easily the DT CF-IPM. It is the only CF-IPM to satisfy the three selection criteria. It has over 90% overall accuracy (and an AUC value of 0.94), less than 5% Type I error, and high transparency. The model (i.e. DT CF-IPM) also made a judicious use of both the quantitative and qualitative variables (see Figure 9.1). RF and AB CF-IPMs could have easily been the joint best choices given their non-existent Type I error and extreme accuracies regarding AUC and overall accuracy. Their opacity, however, implies that they lack a feature (i.e. transparency), or cannot satisfy a criterion, that is of high importance to construction firm’s owners. They hence fall behind DT in selecting the overall best model. The DT model is re-presented here in Figure 9.1. Table 9.2 shows the criteria each of the 13 CF-IPMs satisfies.

Table 9.2: Selecting the model of choice for the research

S/N	Tool/Algorithm	Selection criteria			Selected model of choice
		High Accuracy (>90%)	Low Type I error (<10%)	Transparency	
1.	**Big Data LDA**	✕	✕	✓	
2.	**Big Data QDA**	✕	✕	✓	
3.	**Big Data LR**	✕	✕	✓	
4.	**Big Data NB**	✓	✓	✕	
5.	**Big Data SVM**	✕	✕	✕	
6.	**Big Data KNN**	✓	✓	✕	
7.	**Big Data ANN**	✕	✕	✕	
8.	**Big Data DT**	✓	✓	✓	●
9.	**Big Data RF**	✓	✓	✕	
10.	**Big Data BM**	✓	✓	✕	
11.	**Big Data AB**	✓	✓	✕	
12.	**Big Data PRL**	✓	✓	✕	
13.	**Big Data KHN**	✕	✕	✕	

The DT model is quite easy to read. The most important factors affecting the (in)solvency of construction firms, as given by the DT CF-IPM in Figure 9.1 are R3, R5, R6, Q3, Q4, Q8, Q13. Having produced four qualitative variables and three quantitative variables, the model clearly displays the importance of qualitative variables to predicting insolvency of construction firms. These seven variables, produced by the DT CF-IPM, will be discussed in the next section.

Interpreting the model, the structure indicates R6 to be the most important variable, but the variable cannot independently give a final prediction. The R3 and Q13 variables are the joint second variables and are totally dependent R6. The Q13 variable is, however, able to give a final prediction. So a construction firm that possesses an R9 value greater than 1.90 simply needs to use the Q13 variable to predict its status (i.e. failing or healthy). If a construction firm possesses an R9 value less than 1.90, then it goes through R3 and continues down the chain until a final prediction is made in the line of Q3 and Q4, or Q8 andR5. The rest of the tree (i.e. the DT CF-IPM) can be interpreted as explained here.

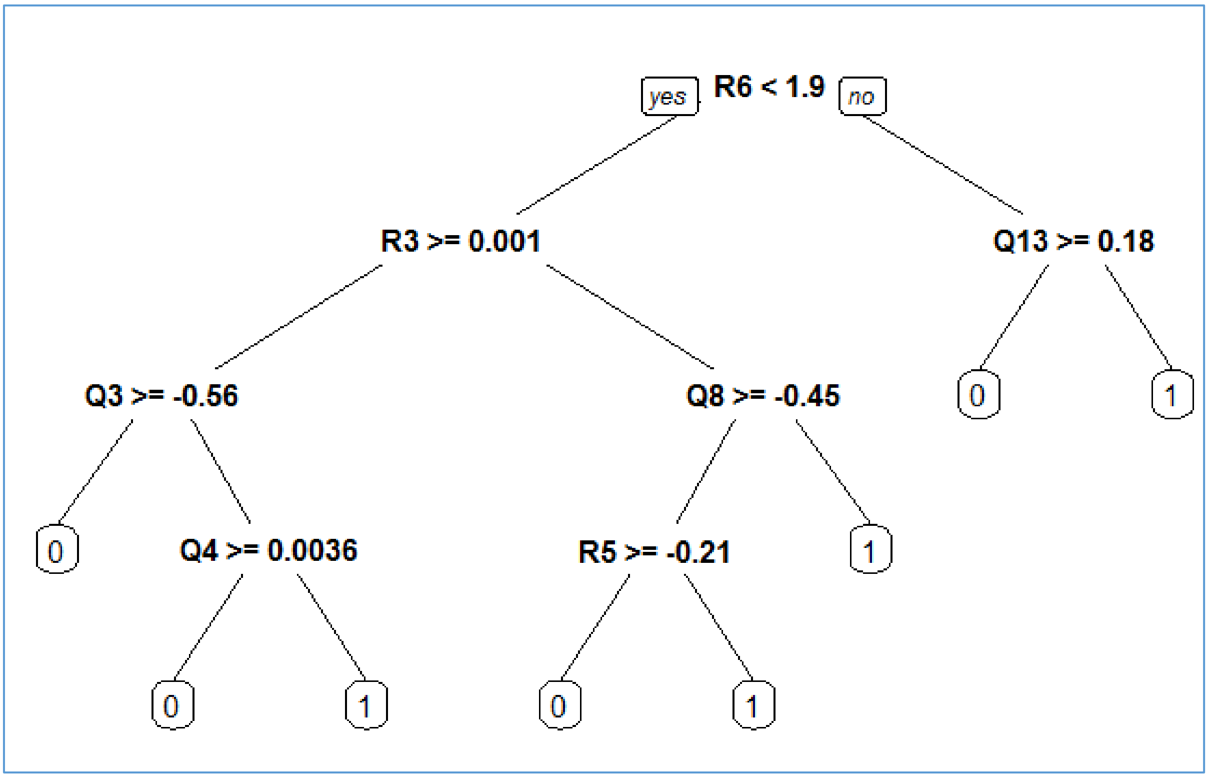


Figure 9.1: The DT model

9.3 Factors affecting (in)solvency of construction firms

The factors affecting the (in)solvency (i.e. failure or survival) of a construction firm, based on the result of the selected CF-IPM are discussed in this section. Factors in this section refer to the variables produced by the DT CF-IPM. The factors are listed in Table 9.3. The qualitative factors (variables) are discussed based on some of their offspring factors gotten from factor analysis (see Table 7.5)

Table 9.3: Factors affecting the (in)solvency of a construction firm

Variable category	Serial number	Variable name
Quantitative variables	R3	Return on Total Assets (%)
	R5	Liquidity ratio
	R6	Solvency ratio (Asset-based) (%)
	Q3	Top management characteristics 1
Qualitative variables	Q4	Strategic issues and external relations
	Q8	Finance and conflict related issues
	Q13	Industry contract/project knowledge

9.3.1 Quantitative factors

Return on Total Assets: This is a measure of profitability of a firm. Profitability ratios are used to measure the entire accomplishment, or returns, which a construction firm's management has realised(Edum-Fotwe, Price and Thorpe, 1996). Profitability is obviously an important factor for the solvency of any business including construction firms; after all, most businesses are started to make profit. In Beaver's (1966) pioneering work, the profitability ratio was the second most important ratio for insolvency prediction after cash flow ratios. The profitability factor's contribution to insolvency prediction is very high (Altman 1968; Taffler 1982; Horta and Camanho 2013) hence it has featured vehemently in predictions models. According to Dimitras, Zanakis, and Zopounidis (1996), profitability reveals the viability of a firm.

A review of the CI literature clearly reveals that profitability is one of the most important financial factors to be considered if an effective IPM is to be built for construction companies (e.g. Bal, Cheung, and Wu, 2013; Chen, 2012; Edum-Fotwe *et al.*, 1996; Horta and Camanho, 2013; Horta, Camanho, and Moreira da Costa, 2012; B. R. Kangari and Farid, 1992; Kapliński, 2008; Mason and Harris, 1979; Russell and Zhai, 1996). According to Arditi *et al.* (2000), the single most common budgetary factor that has led to the failure of construction firms is insufficient profit. This is because of the extremely aggressive bidding with far from accurate estimates and the 'one-off and custom- made production' systems that are synonymous with the construction industry (CI). Horta *et al.* (2012) noted that innovation is key to the profitability of a construction firm.

Ideally, the higher the profitability ratio of a construction firm, the more solvent the firm is taken to be. However, developers using the multi-discriminant analysis (MDA) statistical tool to develop a CF-IPM need to be careful as the tool sometimes wrongfully assign a negative sign to the profitability ratio (see Abidali and Harris, 1995; Mason and Harris, 1979). This problem is commonly known as the counter-intuitive sign problem and has been suggested to be a result of highly collinear variables

Liquidity ratio: Liquidity is generally concerned with a construction firm's "ability to meet its short-term commitments" (Edum-Fotwe *et al.* 1996: p.190). It has to do with how quickly a firm can turn its assets into liquid cash. The liquidity factor is a very important financial factor required to measure the solvency of construction firms and is consequently common in CF-IPM studies (Altman 1968; Kangari and Farid 1992; Koksal and Arditi 2004; Kapliński 2008; Horta *et al.* 2012; Bal *et al.* 2013; Horta and Camanho 2013). Liquidity factor's importance is evidenced in the interests many stakeholders like material suppliers, site employees and staff in general have in it. The interest is because liquidity indicates to what extent a company can meet its commitments without 'liquidating the non-liquid assets' (Horta *et al.* 2012; Horta and Camanho 2013); inability to cover such liabilities which generally leads to insolvency. Generally, the more liquid a construction firm is, the healthier (Edum-Fotwe, Price and Thorpe, 1996).

It has however been noticed that liquidity ratios are not stable over a long period and are thus not effective for early warning systems that are developed to predict potential failure from over four years before actual failure [Bilderbeek (1977) as cited by Altman (1984)]. Liquidity might be poor for early prediction but is very important for construction firms as cash availability is vital for construction projects.

Evaluation of liquidity depends on how organisational assets and liabilities are classified (Saleem, Ur and Rehman, 2011); such classification can greatly affect the insolvency prediction of a certain construction firm. Imagine an asset reclassification that allows more assets of a construction firm to be classified as liquid! The firm suddenly becomes more solvent without any changes at all.

Solvency ratio: This is the same as leverage ratio and, from its name, is obviously important to insolvency prediction. Leverage, as opposed to equity, refers to the amount of borrowed money that is used to finance a firm. According to McGurr and DeVaney (1998) and Dimitras *et al.*, (1996), solvency/leverage ratios are the most vital discriminants and vary by industry characteristics (Saleem, Ur and Rehman, 2011). They are deemed the most powerful indicators of insolvency prediction for construction firms (Edum-Fotwe *et al.* 1996). Typically, the lesser the leverage (ratios) of a firm, the better, although there is no maximum value that depicts automatic insolvency of a construction firm

As opposed to liquidity, leverage ratios measure long-term solvency and thus contribute greatly to early warning systems for the construction firms (Horta and Camanho 2013). Because construction work is normally paid for only when they have been completed, usually on a monthly basis or longer when delayed, construction contractors are exposed to high debt (leverage) normally acquired to pay subcontractors and suppliers. TRhese debts make construction firms more susceptible to failure from leverage (Arditi, Koksal and Kale, 2000).

9.3.2 Qualitative factors

Top management (TM) characteristics 1: This includes TM flexibility (i.e. non-autocratic), creativity or innovation, support to staff, competence, motivation, among others (see Table 7.5 for offspring factors). Autocracy leads the race in this class and is synonymous with an executive with too much power or a person holding multiple executive positions, all which cause failure of construction firms. A very powerful dual-position CEO/chairman, nullifying the all-important managerial power of the chairman being able to sack a defective CEO, is a common feature of failed construction firms (Abidali and Harris 1995; Hall 1994). On the reverse, more flexible executives with each holding a single role will bring about a balanced power, ensuring there are checks and balances to cut any excesses from any angle. A more balanced executive system will help improve the solvency of a construction firm as each TM delegation can perform its duty under the supervision of another. For example, a construction firm with a balanced board without bias for particular personnel will ensure construction project managers are being monitored and cannot make unsupervised/unjustified decisions (Pearce and Zahra 1992). In such a system, construction project managers will be aware that their jobs can be on the line if a project is deviating from plan, and will thus put in the right effort to ensure projects go according to plan

The indecisiveness and inflexibility of a construction company’s owner/TM lead to not realising the available opportunities and threats to the business. When business is slow, a construction firm specialised in pile foundation installation, for example, should be able to decisively identify opportunities of excavation projects and use its excavators for executing such projects. There is a need for the owner/TM of construction companies to be always alert to alternative opportunities. Failure to do this will lead to no reliable strategy formed to avert or take threats or opportunities respectively.

Demotivation of construction project managers, even where there is no autocracy or extremely powerful CEO, can be a root cause of insolvency of a firm (Abidali and Harris 1995). The same thing applies to the general workforce of a construction firm where the absence of experienced or motivated project personnel can cause poor project execution and lead to company failure(Hall, 1994).

Strategic issues and external relations: The offspring factors here include conflict /litigation/legal issues, high immigration and delay of payments to subcontractors, among others. The construction industry's high tendency of having poorly worded contracts, increased project cost and duration, poor quality project delivered, among others, make conflict synonymous with construction projects (Jaffar, Tharim and Shuib, 2011). Conflict can thus stand as another word for construction projects; its root causes, according to various studies, are numerous (Kumaraswamy, 1997; Jaffar, Tharim and Shuib, 2011). Most construction conflicts usually result in litigation. A firm with continuous problems of litigations and legal costs as well as fines and damages payments will probably fail in the long run (Mitkus and Mitkus 2014)

Economic recession is probably the most severe market force for construction firms insolvency as identified by interview respondents and in other studies (e.g. Arditi *et al.*, 2000; R. Kangari, 1988; Kapliński, 2008; Ng, Wong, and Zhang, 2011). For example, an interview respondent said,

> "*I understand property investment and* ***always buy houses and lands and sell them later.*** *Brother, this brings more money to do the building [i.e. construction]. The stupid* ***problem with economy [recession] caused all my property to go down [i.e. devalue]****. Brother, why is America problem our problem (hisses)*"

Although economic recession does not happen too frequently, its effect, when it does, can be devastating. Virtually everyone in the country is hit somehow and plans for new build, renovations, expansions, among others, are widely cancelled if they are not absolutely necessary. The result is a higher contractor/projects ratio. Bigger construction firms that lose out on the few bids available in their class suddenly become hawkish and encroach on the projects small construction firms would normally take, putting them in more danger of shutting down. This makes firms focus a lot on their competitors as a means of survival. A small firm, for example, will do anything to know how much its competitor has put in for a bid and will want to beat it all cost, even if it is at a minor loss, with the hope of repeat business and starving the competitor to death. One potential solution to the economic recession effect is to continuously seek proper information (Marcella and Illingworth 2012) as there are usually hints about such events (economic recession), then create a strategic plan (Mintzberg's perspective). With this, owners can proactively take decisive actions e.g. closing firm down early before any losses in the worst case.

On 'immigration', the challenge highlighted by interview respondents was the open EU border that allows people from other EU countries to work unrestrictedly in the UK. The major complaint was that some probably unregistered skilled workers were able to take especially small renovation and refurbishment jobs for unbelievably low prices. On the other hand, cheap construction labour immigrants favour big construction firms as employing or contracting them helps reduce their cost/wages (Beaverstock and Hall 2012). The immigration problem is somewhat similar to that of too many new firms springing up as they both represent threat of new entrants (see section 9.4).

Delayed payment to subcontractors is highly related to the leverage and liquidity levels of a construction firm. As noted earlier, because construction work is normally paid for only when they have been completed, usually on a monthly basis or longer when delayed, construction contractors are exposed to high debt (leverage) typically acquired to pay subcontractors (Arditi, Koksal and Kale, 2000). If a contractor reaches its debt limit and consequently make very late payments to subcontractor/suppliers, there will be a distrust from the subcontractors/suppliers and future collaborations can be highly bumpy. In the worst cases, subcontractors/suppliers could decline to offer services or request for payment before or during service execution. Both cases can make things more difficult for a construction firm and eventually lead to its insolvency.

Finance and conflict related issues: The offspring factors here include the percentage of firm's earnings used in construction operations, cash flow and submission of very low bids because of fierce competition, among others. A construction firm is substantially reliant upon the success of its construction projects hence for a construction firm to be more solvent, a reasonable size of the firm's cash flow must be employed in operations with a reduced cash in investment (Arditi, Koksal and Kale, 2000; Chen, 2012). This is because of the cash flow conditions of firms in the CI where:

- ❖ Client only pays for completed work that has been financed by the firm, usually on a monthly basis
- ❖ A percentage (normally 10%) of payment is held back by client for potential omissions and defects

It is thus almost impossible for firms to recover expenses, not to mention make profit, before completion of projects. A robust cash flow plan for operations is thus necessary to avoid extreme leverage, being cash strapped or having a negative cash flow, all of which risk the survival of a construction firm (Kale and Arditi 1999). The challenge is to achieve a positive cash flow from project(s) since a negative cash flow increases risk of its survival

The more successful bids a construction firm gets, the more it grows and the more solvent it becomes; lack of successful bids is tantamount to failure (Bal, Cheung and Wu, 2013). Bidding in an area of expertise ensures a competitive low bid thus a firm must have an, or identify its, area of strength where it is unique over competitors. The importance of competitiveness cannot be over emphasised. However, when the economy is not booming, and the ratio of available projects to construction firms is very low, competitions get extreme and construction firms get to submit unrealistically low bids, leading to terminal losses (Arditi, Koksal and Kale, 2000).

Industry contract/project knowledge: The possibility of a construction firm piling up business knowledge and skills through organisational learning is largely dependent on its age(Arditi, Koksal and Kale, 2000). Such learning over time, and the resulting knowledge and skills, help a construction firm to identify favourable markets (e.g. foundation specialisation, residential housing, road construction, among others) for the resources it possesses; create a positive image; establish the important partnership with construction materials suppliers and subcontractors; build positive relationship with financial institutions and potential clients; easily adapt to latest technologies [e.g. Building information modelling (BIM) software; drones on large construction sites, and so on], among others, (March, 1991), all of which their combined absence can lead to a firm's failure.

The problem of 'collecting receivables' is a big one, especially for small construction, firms (Arditi, Koksal and Kale, 2000). This is because construction firms are known for carrying out services in advance of payment hence a poor debt collection system can be quite detrimental (Arditi, Koksal and Kale, 2000). From the stories of respondents, it appears collecting payment for work done has been a 'pain in the neck' for small construction firms especially. A potential solution might be to check pattern of collections and analyse what has led to quick collection of receivables in the past. The successful patterns can then be retained while ferocious effort is made to dumping elements that have led otherwise.

9.4 Implication to theory

The result clearly indicates the multi-theory basis of construction firm’s insolvency. The theory underpinning the offspring factors of the qualitative factors (i.e. variables) were given in Table 6.6. A careful look at the table shows how the most important factors to construction firms insolvency prediction, as discussed in the prior section, are underpinned by internal related, external related, and combinatorial theories (see chapter two).

For instance, when the economy is in recession like in the case of ‘the 2008 global financial crises’, the ratio of available projects to construction firms is very low. The inevitable failure of some construction firms simply as a result of this poor economic situation will then materialise; this failure is underpinned by the organisation ecology theory. The poor economy, in this case, can be regarded as the environment (ecology) which is naturally picking the firms that will fail or survive.

The immigration problem (‘high immigration levels in UK’) is somewhat similar to that of too many new firms springing up as they both represent threat of new entrants. When there is no barrier to entry (Porter’s theory), as is the case in the construction industry, and anyone or any firm can decide to start construction works, then the market can be easily over flooded with firms, leading to tipped balances, fierce competition and insolvencies (Burtonshaw-Gunn, 2009). Using strategy as ploy (Mintzberg’s 5P’s theory) to distract or deter competitors, for example reporting unregistered workers who avoid tax might increase likelihood of survival. This sort of ploy is quite applicable in this case since the main complaint from the interview respondents is that of skilled Europeans (non-UK) who offer very cheap works because they are unregistered and do not pay taxes or national insurance. The immigration problem could be further viewed as an organisation ecology theory underpinned issue with the mass immigration actively changing the environment to a harsh one for small construction firms that focus on small jobs that can be easily done by a single, or two, skilled person(s). The same environment will be supportive of big construction firms as they get the opportunity to employ skilled Europeans for a cheaper salary.

The submission of very low bids due to fierce competition is very much underpinned by the adaptationist perspective through organisational learning (or experience). This is because the bid submitted by a firm is highly dependent on its experience. This assertion can be supported from various positions. One position is that an inexperienced firm will have less understandings of cost of materials and especially processes and can thus unknowingly turn in an unrealistically low bid. Another is that firms who have continually been unsuccessful in their bids will keep driving the total cost of their subsequent bids down out of desperation to secure a contract. The firm uses its experience from previously submitted bids to continually reduce prices on the items in a bid on a relative basis, partially ignoring loss problems. Finally, a very experienced firm could submit a very low bid simply because it understands the industry well and knows how it will make money from other activities attached to the project but not directly stated in the bid. In the last two cases, the construction firms are simply adapting to the environment of the construction industry in order to survive.

The problem of 'unsuccessful collection of payment for completed works' is a big one as highlighted in the previous section. The problem occurs usually because small firms do not possess the resources (resource based view) required to force clients to pay e.g. powerful lawyers, or the luxury to arrange for a stringent payment process that will ensure non-default. They need to find quick solutions to this common construction industry issue i.e. they need to adapt quickly or they will fail (adaptationist perspective).

The case of less flexible TMT/CEO with high resistance to change normally leads to innovation killing. Such a behaviour of TMT, underpinned by the upper echelon theory, can lead to rejection of the contemporary resources (resource based view) that can increase the competitive advantage of a construction firm. For example, a construction firm's TMT that rejects, or rejected, the adoption of Building Information Modelling (BIM) as a contemporary construction process will now be ineligible for all UK government contracts. This ineligibility will be as a result of the UK government BIM Mandate policy which came into force in April 2016. If a high percentage of such a construction firm is from the government, then it is staring insolvency in the face.

Overall, there is no singular theory that seems to perfectly explain the holistic (in)solvency situation of construction firms as demonstrated in the prior paragraphs. Neither is there is a singular group of theories (external, internal or combinatorial) that does so. The implication of the research on theory is that it clarifies the need to amalgamate and refine various relevant parts of existing theories, or existing group of theories, to fully or almost fully explain the (in)solvency situation of construction firms.

9.4 Chapter summary

The results of the CF-IPMs developed in chapter eight were presented and discussed in this chapter. The best model was selected based on three selection criteria that are important to the aim of the research: accuracy, error type levels and transparency. The DT CF-IPM appeared to be the only model that satisfied the three criteria to very high levels. Although RF and AB CF-IPMs possessed exceptional accuracy and extremely low Type I error, they did not satisfy the important transparency criterion

The quantitative and qualitative factors (i.e. variables) affecting (in)solvency of construction firms as produced by the DT CF-IPM include: Return on Total Assets (R3), Liquidity ratio (R5), Solvency ratio (Asset-based) (R6), Top management characteristics 1 (Q3), Strategic issues and external relations (Q4), Finance and conflict related issues (Q8) and Industry contract/project knowledge (Q13). Each of these factors was discussed in relation to construction firms and the construction industry

The most important factors to construction firms' insolvency prediction, as produced by the DT CF-IPM, are underpinned by the internal related, external related, and combinatorial theories, discussed in chapter two. Overall, no singular theory was found to flawlessly explain the holistic (in)solvency situation of construction firms. The result supports the multi-theory perspective to the insolvency of construction firms. The implication of the research on theory is that it clarifies the need to amalgamate and refine various sections of existing theories, or existing group of theories, to fully or almost fully explain the (in)solvency situation of construction firms.

Chapter ten contains a comprehensive conclusion to the research and the contributions of the study. The conclusions are based on the five objectives of the study

CHAPTER TEN

10.0 CONCLUSIONS

10.1 Chapter introduction

This chapter is a conclusion of the research. Section 10.2 is a presentation of the summary of findings by study objectives and the conclusions made based on the findings. Subsections 10.2.1 through 10.2.4 are presentations of findings' summary and conclusions for objectives one through four of the research respectively. Section 10.3 is a highlight of the contributions of the research. While the contributions to academic knowledge are given in subsection 10.3.1, contributions to industry are given in subsection 10.3.2. Sections 10.4 and 10.5 are presentations of limitations of the research and future research opportunities respectively. Section 10.6 is a summary of the chapter

10.2 Summary of findings

10.2.1 Objective One: To identify qualitative variables that contribute to solvency/insolvency of construction firms through literature review and fieldwork.

The identification of qualitative variables was through systematic reviews of past CF-IPM and construction firm failure studies, as well as through interviews with past owners and directors of failed and existing construction firms. The result from these processes were used to formulates questionnaires with ratings that eventually represented the qualitative variables. The qualitative variables discovered covered areas including management/owner characteristics, firm characteristics, among others.

It can be concluded that the importance of the qualitative variables to the CF-IPMs developed cannot be overemphasized since the overall best CF-IPM (i.e. DT CF-IPM) used a total number of seven variables, four of which were qualitative. Further, four (Q8, Q4, Q13 and Q10, in that order) out of the ten leading variables of the most accurate models [i.e. random forest (RF) and adaptive boosting (AB) CF-IPMs] were also qualitative. Although RF and AB, which are ensemble classifiers and produced CF-IPMs with extreme accuracy, are not transparent, their likely set of most important variables were gotten from the CForest variable selection technique (see Figure 8.7 for variable ranking by CForest technique). This is because CForest is also an ensemble technique.

Since the variables used by the highly accurate CF-IPMs and the overall best CF-IPM (i.e. DT CF-IPM) included a reasonable number of qualitative variables, it can be assuredly concluded that qualitative variables stand as essential for the development of CF-IPMs. This did not come as a big surprise since many studies, as highlighted in chapter three, have already established that the most valid insolvency prediction model (IPM) for construction firms cannot be developed without qualitative variables. The onerous task of obtaining data for these variables, compared to getting financial ratios from financial statements, however, explains why they have been left out of most studies.

10.2.2 Objective Two: To identify the quantitative variables (i.e. financial ratios) that are commonly reported by large, medium, small and micro construction firms.

Information on construction firms' financial ratios, which represented quantitative variables, were gotten from FAME (Forecasting Analysis and Modelling Environment) Bureau Van Dijk UK financial database. A thorough study of the financial statements of numerous large and MSM construction firms hosted on the database revealed that there are some financial ratios commonly reported by all sizes of construction firms. Out of the 29 standard financial ratios provided by the database, 11 were recognized to be commonly reported by all categories of construction firms (i.e. large, MSM, failed and existing).

It can be concluded that the significance of the quantitative variables to the CF-IPMs developed cannot be overstated because the overall best CF-IPM (i.e. DT CF-IPM) utilised three quantitative variables as part of its seven variables for development. Further, quantitative variables constituted more than five (R10, R8 R6, R1, R2 and R4, in that order) of the ten significant variables utilized by the highly accurate CF-IPMs (see Figure 8.7 for variable ranking by CForest technique). In fact, some CF-IPMs like the ones developed with statistical tools [i.e. LDA, QDA and LR) recognised only financial variables as being important

Considering the accuracy levels achieved by the CF-IPMs developed, it is concluded that the possiblility of developing one robust and valid CF-IPM for all sizes (i.e. large and MSM) of construction firms is high and real. This is against the popular perception that only firms of similar sizes should be used as a sample for a CF-IPM. In fact, this combination of sizes, with more MSM firms in the sample, is more representative of the size distribution in the construction industry hence, the CF-IPMs developed here can be concluded to be more representative and are more valid.

Revelation of the importance of the quantitative variables is quite expected because financial ratios have long been established to be very good predictors of insolvency for all types of firms including construction. Although their exclusive use for CF-IPMs is questionable, their general validity was never in doubt and that has been reiterated by the results here.

10.2.3 Objective Three: To select the best combination of quantitative and qualitative for the CF-IPM.

To achieve the third objective, 11 advanced techniques were used to select the best combination of quantitative and qualitative variables. The results from the techniques were different, explaining why there is a disparity in the variables selected in different past studies that used differing variable selection techniques. A voting system was used to select the final variables for model development. The high-performance levels exhibited by most of the CF-IPMs makes it concludable that the final combination of variables selected to develop them were clearly the best.

Since the overall best and most accurate CF-IPMs utilised both the qualitative and quantitative variables in achieving their results, it can be concluded that the best IPMs for construction firms can only be developed using a combination of qualitative and quantitative variables as advocated by a number of studies (see chapter three on variables). The following facts further reinforce this conclusion:

(i) The only past CF-IPM study (Hall, 1994) that utilised only qualitative variables failed in its attempt

(ii) that all the CF-IPMs studies that utilised mainly quantitative variables (i.e. the statistical tools' CF-IPMs) all performed sub-optimally

(iii) that no variable selection technique selected all variables of a particular type, quantitative or qualitative, before selecting the other.

10.2.4 Objective Four: To use advanced well-tuned AI tools and the best contemporary methods to ensure dependability of the CF-IPM

Thirteen tools were used to build the CF-IPMs including ten well-tuned AI tools. This caused some computation intensity challenges which were dealt with and concluded on in subsection 10.2.5. Results showed that the more popular AI tools, i.e. support vector machine (SVM) and artificial neural network (ANN), underperformed along with statistical tools, when compared to the more advanced AI tools. Although high performances from ANN and SVM are well documented in IPM studies, those performances were based on exclusive use of financial ratios as variables. In addition, the counterpart poor statistical tools' CF-IPMs in the proposed solution clearly did not attach any importance to the qualitative variables. It can thus be concluded that contemporary AI tools (e.g. RF, AB, among others) are best for developing CF-IPMs since only they adequately handled qualitative variables which are instrumental to the validity of CF-IPMs. It is also concluded CF-IPM sudies need to start exploring more AI tools since over 20 of them are available. Overall, ensemble classifiers (i.e. RF and AB) produced the most accurate models but have no transparency

With no CF-IPM performing overly well on predicting one class (i.e. failed or existing class) over the other, it can be concluded that the use of data with nearly equal dispersion brought about unbiased CF-IPMs. The use of contemporary measures like receiver operator characteristic (ROC) curve and area under curve (AUC) accuracy values aided successful comparison of closely performing CF-IPMs from two tools: RF and AB (see section 9.2). The model validation/testing process ensured the CF-IPMs were correctly tested while error type consideration ensured that models were evaluated based on the aim of the study rather than just overall accuracy. For instance, despite having a slightly higher overall accuracy, the AB CF-IPM was ranked below the RF CF-IPM because while the RF CF-IPM had zero percent Type I error (the costlier error), the AB CF-IPM had 0.65%. It can thus be concluded that the use of contemporary methods was essential to building high performing CF-IPMS, and selection of the best model.

10.2.5 Objective Five: To solve the high computation intensity problem of large data and tuned AI tools by using Big Data Analytics to develop the CF-IPM.

The data used to build the CF-IPMs contained 14 variables of more than a thousand sample construction firms. A CF-IPM development attempt with highly tuned ANN, using this data on a normal computer, had to be aborted after two days without success. It can be recollected that Du Jardin's (2010) attempt with 500 firms' data took a duration of five days with 30 computers. Despite these facts, the same data of over a thousand construction firms was 'eventually' used to develop highly tuned sophisticated CF-IPMs in 'seconds', simply by using Big Data Analytics. It can thus be concluded that a highly tuned, sophisticated and very reliable CF-IPM with massive data can now be developed easily and quickly by using a contemporary analytics technology: Big Data Analytics. It can also be concluded that Big Data analytics is useful for, applicable to, and can help, construction firms and the construction industry.

10.3 Contributions of study

10.3.1 Contribution of Study to Academic Knowledge

One of the greatest contributions of the research is that it has demonstrated how Big Data Analytics can be used to develop CF-IPM. It has been successfully shown in the research that a very large amount of data can be used with highly tuned AI tools to develop IPMs for construction firms in seconds rather than in days.

Analysing narratives of respondents, including top management team members and owners of large and MSM construction firms, the research has contributed a number of qualitative variables for developing IPMs for construction firms to the CF-IPM literature. The study has also shown how quantitative and qualitative variables can be combined to develop IPMs for construction firms.

Another contribution of the research is the establishment of the fact the LDA, QDA, LR, ANN and SVM are unfit for developing CF-IPMs if the all-important step of combining qualitative and quantitative variables is to be taken. This is a very important contribution since most of these tools are quite common with CF-IPM studies. It must, however, be emphasised that this conclusion is only valid when qualitative variables are utilised as ANN and SVM are known to perform very well when using only quantitative variables.

Further, the research has exposed this area of research to more powerful AI tools such as RF, AB and Bart machine, among others. This will enlighten other authors and expose them to the fact that there are many powerful AI tools out there other than the ones that are commonly used in CF-IPM or general IPM studies.

10.3.2 Contribution of Study to Practice

There are two major contributions of the research to the industry. The first is the capability of the CF-IPMs developed in the research to identify potential failure relatively early. This is because of the qualitative variables involved in their development. It is well known that financial ratios only reflect failure of a firm at the point of death of the firm. In fact, financial ratios are a late reflection of earlier managerial decisions which are only measured with qualitative variables. Qualitative variables thus improve early predictive capabilities of CF-IPMs.

The second major contribution is the CF-IPMs capability to carry out predictions for large and MSM firms, as against previous CF-IPMs which focus on just larger, and maybe medium sized, firms because of the completeness of their financial statements. This is very important for the construction industry as it consists mainly of micro and small construction firms compared to a relatively few medium and large construction firms. The CF-IPMs also used qualitative variables, giving a better chance of usage to small and micro construction firms who may have incomplete financial statements.

10.4 Limitations of study

Every research comes with some form of limitations and the research is not an exemption. The chief limitation of the research is the inability to recognise construction firms that have falsely declared bankruptcy or falsified insolvency conditions. It is almost practically impossible to recognise such construction firms because perpetrators normally understand bankruptcy/insolvency laws well enough to lawfully declare bankruptcy or be declared insolvent. The trouble with the potential inclusion of such construction firms in the data for the research is that many of its untampered with features (i.e. variables) could easily represent those of an existing/healthy construction firm. Its placement among insolvent construction firms would thus be wrong and could affect accuracy levels of the CF-IPMs.

A less impacting limitation of the research is the oversampling of the 531 questionnaire data to achieve another 531 in order to have data of 1062 firms. There is no doubt that having data from 1062 firms directly would be more realistic and better. However, high-level oversampling techniques, like the synthetic minority over-sampling technique (SMOTE) algorithm employed in the research, have been proven over the years to be very effective. Further, the fact that the CF-IPMs had very high accuracy levels predicting classes of real and oversampled firms is a testament to the effectiveness of the technique. In addition, the oversample questionnaire data was paired with real financial data, thereby improving legitimacy of each oversampled construction firm data.

10.5 Future research opportunities

Based on the process and output of the research, one area that future research should look into is the possibility of developing a form of qualitative variables document which will serve as equivalent to financial statement. This document can then be completed annually by construction firms making IPMs for construction firms much easier to develop.

The research also successfully proved the multiple theory basis of insolvency of construction firms. It gives the opportunity for future studies to create a holistic theory, from the established set of theories, underpinning the failure of construction firms.

10.6 Chapter summary

It is concluded in this chapter that the utilisation of both quantitative and qualitative variables by the best performing CF-IPMs shows the extreme importance of combining both types of variables to insolvency prediction of construction firms. The high-level performances achieved by many of the CF-IPMs is proof that it is very possible to build one robust and valid model for all category sizes (i.e. large and MSM) of construction firms. The inclusion of more MSM firms in the research’s sample makes it more representative of the construction industry. From the results, it can further be concluded that LDA, QDA, LR, ANN and SVM are all unfit for developing IPMs for construction firms since they cannot handle qualitative variables which are instrumental to the validity of IPMs for construction firms.

The research has contributed some qualitative variables to the CF-IPM research area and has clearly established that there are many other high-performing AI tools (e.g. RF and Bart machine among others) that are not yet being explored in the CF-IPM research area. The contribution to industry is the early predictive capability of the CF-IPMs and their industry-wide usefulness in terms of being relevant to all sizes of construction firms

The chief limitation of the research is the inability to recognise construction firms that have falsely declared bankruptcy or falsified insolvency. A less impacting limitation of the research is the use of oversampling of the 531 questionnaire data to achieve another 531 in order to have data of 1062 firms. Future research should look into is the possibility of developing a form of qualitative variables document which will serve as equivalent to financial statement. They should also seek to create a holistic theory, from the established set of theories, underpinning the failure of construction firms.

REFERENCES

Abatecola, G. and Cristofaro, M. (2016) Upper Echelons and Executive Profiles in the Construction Value Chain: Evidence from Italy. *Project Management Journal*. 47 (1), pp. 13–26.

Abidali, A.F. and Harris, F. (1995) A methodology for predicting company failure in the construction industry. *Construction Management and Economics*. 13 (3), pp. 189–196.

Acid, S., de Campos, L.M. and Fernandez, M. (2011) Minimum redundancy maximum relevancy versus score-based methods for learning Markov boundaries. In: *2011 11th International Conference on Intelligent Systems Design and Applications*. November 2011 Cordoba: IEEE. pp. 619–623.

Adams, E. and Simon, H.A. (1962) Models of Man, Social and Rational: Mathematical Essays on Rational Human Behavior in a Social Setting. *The Journal of Philosophy*. 59 (7), pp. 177.

Adnan Aziz, M. and Dar, H.A. (2006) Predicting corporate bankruptcy: where we stand? *Corporate Governance: The international journal of business in society*. 6 (1), pp. 18–33.

Agarwal, V. and Taffler, R. (2008) Comparing the performance of market-based and accounting-based bankruptcy prediction models. *Journal of Banking and Finance*. 32 (8), pp. 1541–1551.

Agneeswaran, V.S. (2014) *Big Data Analytics: Evolution of Machine Learning Realizations*. New Jersey: Pearson Education.

Ahn, B.S., Cho, S.S. and Kim, C.Y. (2000) The integrated methodology of rough set theory and artificial neural network for business failure prediction. *Expert Systems With Applications*. 18 (2), pp. 65–74.

Ali Khan, S., Hussain, A., Basit, A. and Akram, S. (2014) Kruskal-Wallis-based computationally efficient feature selection for face recognition. *The Scientific WorldJournal*. 2014 pp. 672630.

Altman, E. (1968) Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*. 23 (4), pp. 589–609.

Altman, E. (1984) The success of business failure prediction models: An international survey. *Journal of Banking and Finance*. 8 pp. 171–198.

Altman, E.I. (1993) *Corporate Financial Distress And Bankruptcy : A Complete Guide To Predicting andAmp; Avoiding Distress And Profiting From Bankruptcy*. New Jersey: Wiley.

Altman, E.I., Marco, G. and Varetto, F. (1994) Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience). *Journal of Banking and Finance*. 18 (3), pp. 505–529.

Altman E.I., Haldeman, R.G., N. (1977) ZETA Tm* ANALYSIS A new model to identify bankruptcy risk of corporations. *Journal of Banking and Finance*. 1 (1), pp. 29–54.

Alvesson, M. and Sköldberg, K. (2000) *Reflexive methodology : new vistas for qualitative research*. London: SAGE.

Amburgey, T.L. and Rao, H. (1996) Organizational ecology: past, present, and future directions. *Academy of Management Journal*. 39 (5), pp. 1265–1286.

De Andrés, J., Sánchez-Lasheras, F., Lorca, P. and Juez, F.J.D.C. (2011) A hybrid device of self organizing maps (SOM) and multivariate adaptive regression splines (MARS) for the forecasting of firms' bankruptcy. *Accounting and Management Information Systems*. 10 (3), pp. 351–374.

Andrews, K.R. (1987) *Concept of Corporate Strategy.* Toronto: Richard D Irwin.

Appiah, K.O., Chizema, A. and Arthur, J. (2015) Predicting corporate failure: a systematic literature review of methodological issues. *International Journal of Law and Management*. 57 (5), pp. 461–485.

Arditi, D., Koksal, A. and Kale, S. (2000) Business failures in the construction industry. *Engineering, Construction and Architectural Management*. 7 (2), pp. 120–132.

Argenti, J. (1976) Corporate planning and Corporate Collapse. *Long Range Planning*. 9 (6), pp. 12–17.

Argenti, J. (1980) *Practical Corporate Planning*. Crows Nest: G. Allen and Unwin.

Arslan, G., Tuncan, M., Birgonul, M.T. and Dikmen, I. (2006) E-bidding proposal preparation system for construction projects. *Building and Environment*. 41 (10), pp. 1406–1413.

Atiya, A.. (2001) Bankruptcy prediction for credit risk using neural networks: a survey and new results. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*. 12 (4), pp. 929–935.

Aveyard, H. (2014) Doing A Literature Review In Health And Social Care: A Practical Guide: A Practical Guide *How do I critically appraise the literature?* p.pp. 191.

Awad, A. and Fayek, A.R. (2013) Adaptive Learning of Contractor Default Prediction Model for Surety Bonding. *Journal of Construction Engineering and Management*. 139 (6), pp. 694–704.

Bailey, K.D. (1994) *Methods of Social Research*. New York: Free Press.

Bal, J., Cheung, Y. and Wu, H.-C. (2013) Entropy for Business Failure Prediction: An Improved Prediction Model for the Construction Industry. *Advances in Decision Sciences*. 2013 pp. 1–14.

Balcaen, S. and Ooghe, H. (2006) 35 Years of Studies on Business Failure: an Overview of the Classic Statistical Methodologies and Their Related Problems. *The British Accounting Review*. 38 (1), pp. 63–93.

Barga, R.S., Ekanayake, J. and Lu, W. (2012) Project Daytona: Data analytics as a cloud service. In: *Proceedings - International Conference on Data Engineering*. April 2012 Washington: IEEE. pp. 1317–1320.

Barnett, W.P. and McKendrick, D.G. (2004) Why are Some Organizations More Competitive than Others? Evidence from a Changing Global Market. *Administrative Science Quarterly*. 49 (4), pp. 535–571.

Barney, J.B. (2000) Firm resources and sustained competitive advantage. In: A.C. Baum

Joel and Dobbin Frank (eds.). *Economics Meets Sociology in Strategic Management* 17th edition. Bingley : Emerald Group Publishing Limited. pp. 203–227.

Baruch, Y. and Holtom, B.C. (2008) Survey response rate levels and trends in organizational research. *Human Relations*. 61 (8), pp. 1139–1160.

Baum, J.A.C. and Singh, J. V. (1996) Dynamics of Organizational Responses to Competition. *Social Forces*. 74 (4), pp. 1261–1297.

Baum, J.A.C. and Singh, J. V. (1994) *Evolutionary Dynamics of Organizations*. Oxford: Oxford University Press.

Beaver, W. (1966) Financial ratios as predictors of failure. *Journal of accounting research*. 4 (1966), pp. 71–111.

Beaverstock, J. V. and Hall, S. (2012) Competing for talent: global mobility, immigration and the City of London's labour market. *Cambridge Journal of Regions, Economy and Society*. 5 (2), pp. 271–288.

Becchetti, L. and Sierra, J. (2003) Bankruptcy risk and productive efficiency in manufacturing firms. *Journal of Banking and Finance*. 27 (11), pp. 2099–2120.

Bettany-Saltikov, J. (2012) *How to Do a Systematic Literature Review in Nursing: A Step-By-Step Guide*. Berkshire: Open University Press.

Betts, M. and Ofori, G. (1992) Strategic planning for competitive advantage in construction. *Construction Management and Economics*. 10 (6), pp. 511–532.

Bifet, A., Holmes, G., Kirkby, R. and Pfahringer, B. (2010) MOA: Massive Online Analysis. *Journal of Machine Learning Research*. 11 (May), pp. 1601–1604.

Blaikie, N. (1993) *Approaches to Social Enquiry Polity* p.pp. 243.

Bless, C., Higson-Smith, C. and Kagee, A. (2006) *Fundamentals of Social Research Methods: An African Perspective*. Cape Town: Juta.

Boritz, J.E., Kennedy, D.B. and Albuquerque, A. de M. e (1995) Predicting Corporate Failure Using a Neural Network Approach. *Intelligent Systems in Accounting, Finance* *and Management*. 4 (2), pp. 95–111.

Bourgeois, L.J. (1984) Strategic Management and Determinism. *Academy of Management Review*. 9 (4), pp. 586–596.

Bouwen, R. and Steyaert, C. (1997) elling Stories of Entrepreneurship. Towards a narrative-contextual epistemology for entrepreneurial studies. In: Rik. Donckels and Asko. Miettinen (eds.). *Entrepreneurship and SME Research : on its Way to the Next Millennium*. Aldershot: Ashgate. pp. 47–62.

Boyd, D. and Crawford, K. (2012) Critical questions for big data. *Information, Communication and Society*. 15 (5), pp. 662–679.

Bruderl, J. and Schussler, R. (1990) Organizational Mortality: The Liabilities of Newness and Adolescence. *Administrative Science Quarterly*. 35 (3), pp. 530.

Bryman, A. and Bryman, P. of S.R.A. (2003) *Research Methods and Organization Studies*. London: Unwin Hyman.

Bu, Y., Howe, B., Balazinska, M. and Ernst, M.D. (2010) HaLoop. *Proceedings of the VLDB Endowment*. 3 (1–2), pp. 285–296.

Budayan, C., Dikmen, I. and Talat Birgonul, M. (2013) Investigation of drivers and modes of differentiation in Turkish construction industry. *Engineering, Construction and Architectural Management*. 20 (4), pp. 345–364.

Burrell, G. and Morgan, G. (1979) *Sociological Paradigms and Organisational Analysis - Elements of the Sociology of Corporate Life*. Aldershot: Ashgate.

Burtonshaw-Gunn, S. (2009) *Risk and Financial Management in Construction*. New York: Gower.

Callon, M. (2006) What does it mean to say that economics is performative? *CSI Working Papers Series*.

Carpenter, M.A., Geletkanycz, M.A. and Sanders, W.G. (2004) Upper Echelons Research Revisited: Antecedents, Elements, and Consequences of Top Management Team Composition. *Journal of Management*. 30 (6), pp. 749–778.

Carroll, G.R. and Hannan, M.T. (1995) *Organizations in Industry - Strategy, Structure, and Selection*. Oxford: Oxford University Press.

Chang, A.S.-T. (2001) Defining Cost/Schedule Performance Indices and Their Ranges for Design Projects. *http://dx.doi.org/10.1061/(ASCE)0742-597X(2001)17:2(122)*.

Chawla, N., Bowyer, K. and Hall, L. (2002) SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*. 16 pp. 321–357.

Chen, H.L. (2009) Model for Predicting Financial Performance of Development and Construction Corporations. *Journal of Construction Engineering and Management*. 135 (11), pp. 1190–1200.

Chen, J.-H. (2012) Developing SFNN models to predict financial distress of construction companies. *Expert Systems with Applications*. 39 (1), pp. 823–827.

Cheng, M.-Y. and Hoang, N.-D. (2015) Evaluating Contractor Financial Status Using a Hybrid Fuzzy Instance Based Classifier: Case Study in the Construction Industry. *IEEE Transactions on Engineering Management*. 62 (2), pp. 184–192.

Cheng, M.-Y., Hoang, N.-D., Limanto, L. and Wu, Y.-W. (2014) A novel hybrid intelligent approach for contractor default status prediction. *Knowledge-Based Systems*. 71 pp. 314–321.

Child, D. (2006) *The Essentials of Factor Analysis*. London: Continuum.

Child, J. (1972) Organizational Structure, Environment and Performance: The Role of Strategic Choice. *Sociology*. 6 (1), pp. 1–22.

Chinowsky, P.S. and Meredith, J.E. (2000) Strategic management in construction organizations. In: *In Construction Congress VI: Building Together for a Better Tomorrow in an Increasingly Complex World*. 2000 Orlando: American Society of Civil Engineers.

Chung, K.C., Tan, S.S. and Holdsworth, D.K. (2008) Insolvency Prediction Model Using Multivariate Discriminant Analysis and Artificial Neural Network for the Finance Industry in New Zealand. *International Journal of Business and Management*. 39 (1), pp. 19–28.

Cliff, N. (1988) The eigenvalues-greater-than-one rule and the reliability of components. *Psychological Bulletin*. 103 (2), pp. 276–279.

Creswell, J.W. (2010) *Designing and Conducting Mixed Methods Research*. London: SAGE Publications.

Crossan, M.M., Lane, H.W. and White, R.E. (1999) An organizational learning framework: from intuition to institution. *Academy of Management Review*. 24 (3), pp. 522–537.

Crotty, M. (1998) *The Foundations of Social Research: Meaning and Perspective in The Research Process*. London: Sage Publications.

Cunningham, P. and Delany, S.J. (2007) *k-Nearest Neighbour Classifiers*.

D'hondt, S. (1994) Conversational Realities: Constructing Life through Language *Pragmatics* 4 (1) p.pp. 131–132.

Dainty, A. (2008) Methodological pluralism in construction management research. *Advanced research methods in the built environment*. 1 pp. 1–13.

Dean, J. and Ghemawat, S. (2008) MapReduce. *Communications of the ACM*. 51 (1), pp. 107.

Denning, S. (2005) The Role of Narrative in Organizations. In: J. Brown, S. Denning, K. Groh, and L. Prusack (eds.). *Storytelling in Organizations: Why Storytelling Is Transforming 21st Century Organizations and Management*. Oxford: Elsevier Butterworth-Heinemann. pp. 165–182.

Department for Business Innovation and Skills (2015) *Business Population Estimates for the UK and regions 2015*.

Department for Business Innovation and Skills (2013a) *Industrial Strategy: government and industry in partnership*.

Department for Business Innovation and Skills (2013b) *UK Construction - An Economic Analysis of the Sector, BIS/13/958*.

Dewey, J. (1920) *Reconstruction in philosophy : Dewey, John, 1859-1952 : Free Download* *andamp; Streaming : Internet Archive*. Boston: Beacon Press.

Diebold, F.X. (2012a) On the Origin(s) and Development of the Term 'Big Data' *SSRN Electronic Journal*.

Diebold, F.X. (2012b) On the Origin(s) and Development of the Term 'Big Data' *SSRN Electronic Journal*.

Dikmen, I. and Birgönül, M.T. (2003) Strategic Perspective of Turkish Construction Companies. *Journal of Management in Engineering*. 19 (1), pp. 33–40.

Dimitras, A., Zanakis, S. and Zopounidis, C. (1996) A survey of business failures with an emphasis on prediction methods and industrial applications. *European Journal of Operational …*. (1968), .

Dirickx, Y. and Van Landeghem " ', G. (1994) Statistical Faillure Previsisn Problems. *Tijdschrift voor Economie en Management*. 39 (4), pp. 429–462.

Divsalar, M., Roodsaz, H., Vahdatinia, F., Norouzzadeh, G. and Behrooz, A.H. (2012) A Robust Data-Mining Approach to Bankruptcy Prediction. *Journal of Forecasting*. 31 (6), pp. 504–523.

Dun and Bradstreet Limited (2012) *Global business failures report*.

Van Dyne, L. and LePine, J.A. (1998) Helping and voice extra-role behaviors: evidence of construct and predictive validity. *Academy of Management Journal*. 41 (1), pp. 108–119.

Easterby-Smith, M., Thorpe, R. and Jackson, P. (2008) *Management Research : An Introduction*. London: Sage.

Edmister, R.O. (1972) An Empirical Test of Financial Ratio Analysis for Small Business Failure Prediction. *The Journal of Financial and Quantitative Analysis*. 7 (2), pp. 1477.

Edum-Fotwe, F., Price, A. and Thorpe, A. (1996) A review of financial ratio tools for predicting contractor insolvency. *Construction Management and Economics*. 14 (3), pp. 189–198.

Ekanayake, J., Li, H., Zhang, B., Gunarathne, T., Bae, S.-H., Qiu, J. and Fox, G. (2010) Twister. In: *Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing - HPDC '10*. 2010 New York, New York, USA: ACM Press. pp. 810.

Enshassi, A., Al-Hallaq, K. and Mohamed, S. (2006) Causes of contractor's business failure in developing countries: the case of Palestine. *Journal of Construction in Developing Countries*. 11 (2), pp. 1–14.

Ericson, K. and Pallickara, S. (2013) On the performance of high dimensional data clustering and classification algorithms. *Future Generation Computer Systems*. 29 pp. 1024–1034.

Etzkowitz, H., Webster, A., Gebhardt, C. and Terra, B.R.C. (2000) The future of the university and the university of the future: evolution of ivory tower to entrepreneurial paradigm. *Research Policy*. 29 (2), pp. 313–330.

Everett, J. and Watson, J. (1998) Small Business Failure and External Risk Factors. *Small Business Economics*. 11 (4), pp. 371–390.

Fadel, H. (1977) The predictive power of financial ratios in the British construction industry. *Journal of Business Finance and Accounting*. 4 (3), pp. 339–352.

Fan, W. and Bifet, A. (2013) Mining big data. *ACM SIGKDD Explorations Newsletter*. 14 (2), pp. 1.

Faulin, J. (1999) Multi methodology: The theory and Practice of Combining Management Science Methodologies. *Interfaces*. 29 (2), pp. 135–137.

Field, A.P. (2009) *Discovering Statistics Using SPSS : (And Sex And Drugs And Rock 'N' Roll).* London: SAGE Publications.

Fincham, J.E. (2008) Response rates and responsiveness for surveys, standards, and the Journal. *American journal of pharmaceutical education*. 72 (2), pp. 43.

Fink, A. (2010) *Conducting Research Literature Reviews : From The Internet To Paper*. London: Sage Publishing.

Finkelstein, S. and Hambrick, D.C. (1990) Top-Management-Team Tenure and Organizational Outcomes: The Moderating Role of Managerial Discretion. *Administrative Science Quarterly*. 35 (3), pp. 484.

Foddy, W. (1993) *Constructing Questions for Interviews and Surveys: Theory and Practice in Social Research*. Cambridge: Cambridge University Press.

Freeman, J., Carroll, G.R. and Hannan, M.T. (1983) The Liability of Newness: Age Dependence in Organizational Death Rates. *American Sociological Review*. 48 (5), pp. 692.

Gabriel, Y. and Griffiths, D.S. (2004) Stories in organizational research. In: Catherine Cassell and Gillian Symon (eds.). *Essential Guide to Qualitative Methods in Organizational Research*. London: Sage. pp. 114–126.

George, C. (2002) Why Do Some Companies Thrive While Others Fail? | Stanford Graduate School of Business *Graduate School of Stanford Business*.

George, G. and Mallery, P. (2003) *SPSS for Windows Step by Step: A Simple Guide and Reference*. Massachusetts: Allyn and Bacon.

Giacobbi, P., Poczwardowski, A. and Hager, P. (2005) A Pragmatic Research Philosophy for Applied Sport Psychology. *The Sport Psychologist*.

Global Construction Perspectives and Oxford Economics (2015) *Global Construction 2030. A global forecast for the construction industry to 2030*.

Gov.uk (2014) *Applying to become bankrupt - GOV.UK*. Available from: https://www.gov.uk/bankruptcy [Accessed 26 September 2016].

Gowda Karegowda, A., Manjunath, A.S. and Jayaram, M.A. (2010) Comparative study of attribute selection using gain ratio and correlation based feature selection. *International Journal of Information Technology and Knowledge Management*. 2 (2), pp. 271–277.

Griliches, Z. (1979) Issues in Assessing the Contribution of Research and Development to Productivity Growth. *The Bell Journal of Economics*. 10 (1), pp. 92.

Grünbaum, N.N. (2007) Identification of ambiguity in the case study research typology: what is a unit of analysis? *Qualitative Market Research: An International Journal*. 10 (1), pp. 78–97.

Gu, R., Shen, F. and Huang, Y. (2013) A parallel computing platform for training large scale neural networks. In: *2013 IEEE International Conference on Big Data*. October 2013 Silicon Valley, CA: IEEE. pp. 376–384.

Guba, E.G. (1990) *The Paradigm Dialog*. London: Sage Publications.

Guest, G., Bunce, A. and Johnson, L. (2006) How Many Interviews Are Enough?: An Experiment with Data Saturation and Variability. *Field Methods*. 18 (1), pp. 59–82.

Hadoop (2014) *Welcome to Apache^{TM} Hadoop®!* Available from: http://hadoop.apache.org/ [Accessed 26 February 2015].

Hall, G. (1994) Factors distinguishing survivors from failures amongst small firms in the UK construction sector*. *Journal of Management Studies*.

Hambrick, D.C. and Mason, P.A. (1984) Upper Echelons: The Organization as a Reflection of Its Top Managers. *Academy of Management Review*. 9 (2), pp. 193–206.

Hannan, M.T. and Freeman, J. (1977) The Population Ecology of Organizations The Population Ecology of Organizations'. *Source: American Journal of Sociology*. 82 (5), pp. 929–964.

Harada, N. (2007) Which Firms Exit and Why? An Analysis of Small Firm Exits in Japan. *Small Business Economics*. 29 (4), pp. 401–414.

Harrell, F.E. (2001) *Regression Modeling Strategies*Springer Series in Statistics. New York, NY: Springer New York.

Harris D. (2014) *Gigaom | Apache Mahout, Hadoop's original machine learning project, is moving on from MapReduce*. Available from: https://gigaom.com/2014/03/27/apache-mahout-hadoops-original-machine-learning-project-is-moving-on-from-mapreduce/ [Accessed 26 February 2015].

Heo, J. and Yang, J.Y. (2014) AdaBoost based bankruptcy forecasting of Korean construction companies. *Applied Soft Computing*. 24 pp. 494–499.

Higgins, J.P.T. and Green, S. (2008) Front Matter. In: J.P.T. Higgins and S. Green (eds.). *Cochrane Handbook for Systematic Reviews of Interventions*. Chichester, UK: John Wiley and Sons, Ltd.

Hill, J. and McGowan, P. (1999) Small business and enterprise development: questions about research methodology. *International Journal of Entrepreneurial Behavior and Research*. 5 (1), pp. 5–18.

Hitzler, E.P. and Janowicz, K. (2013) Linked Data, Big Data, and the 4th Paradigm. *Semantic Web,*. 4 (3), pp. 233–235.

Hodgson, M. (2013) *Insolvency figures show 23% of failures come from Construction Industry - Tremark*. Available from: http://www.tremark.co.uk/177-insolvency-figures-show-23-of-failures-come-from-construction-industry/ [Accessed 5 January 2016].

Holloway, I. (1997) *Basic Concepts for Qualitative Research*. Carlton Victoria: Blackwell Science.

Horta, I.M. and Camanho, a. S. (2013) Company failure prediction in the construction industry. *Expert Systems with Applications*. 40 (16), pp. 6253–6257.

Horta, I.M., Camanho, a. S. and Moreira da Costa, J. (2012) Performance assessment of construction companies: A study of factors promoting financial soundness and innovation in the industry. *International Journal of Production Economics*. 137 (1), pp. 84–93.

Horwath, J. (1994) Book reviews: Organizational Change and Redesign: Ideas and Insights for Improving Performance. *Adminsitrative Science Quarterly*. (December), pp. 2344–2345.

Hoshmand, L.T. (2003) Can Lessons of History and Logical Analysis Ensure Progress in Psychological Science? *Theory and Psychology*. 13 (1), pp. 39–44.

Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X. (2013) *Applied logistic regression.* New Jersey: Joh Wiley and Sons.

Huang, S.-C., Tang, Y.-C., Lee, C.-W. and Chang, M.-J. (2012) Kernel local Fisher discriminant analysis based manifold-regularized SVM model for financial distress predictions. *Expert Systems with Applications*. 39 (3), pp. 3855–3861.

Huang, Y. (2009) Prediction of contractor default probability using structural models of credit risk: an empirical investigation. *Construction Management and Economics*. 27 (6), pp. 581–596.

Hycner, R.H. (1985) Some guidelines for the phenomenological analysis of interview data. *Human Studies*. 8 (3), pp. 279–303.

Jaafar, M. and Abdul-Aziz, A.-R. (2005) Resource-Based View and Critical Success Factors: A Study on Small and Medium Sized Contracting Enterprises (SMCEs) in Malaysia. *International Journal of Construction Management*. 5 (2), pp. 61–77.

Jackson, R.H.G. and Wood, A. (2013) The performance of insolvency prediction and credit risk models in the UK: A comparative study. *The British Accounting Review*. 45 (3), pp. 183–202.

Jacobs, A. and Adam (2009) The pathologies of big data. *Communications of the ACM*. 52 (8), pp. 36.

Jaffar, N., Tharim, A.H.A. and Shuib, M.N. (2011) Factors of Conflict in Construction Industry: A Literature Review. *Procedia Engineering*. 20 pp. 193–202.

James, W. (1995) *Pragmatism.* New York: Dover.

Jannadi, M.O. (1997) Reasons for Construction Business Failures in Saudi Arabia. *Project Management Journal*. 28 (2), pp. 185–200.

du Jardin, P. (2015) Bankruptcy prediction using terminal failure processes. *European Journal of Operational Research*. 242 (1), pp. 286–303.

Du Jardin, P. (2010) Predicting bankruptcy using neural networks and other classification methods: The influence of variable selection techniques on model accuracy. *Neurocomputing*. 73 (10), pp. 2047–2060.

Du Jardin, P. and Séverin, E. (2011) Predicting corporate bankruptcy using a self-organizing map: An empirical study to improve the forecasting horizon of a financial failure model. *Decision Support Systems*. 51 (3), pp. 701–711.

Jaunzens, D. (2001) *Influencing small businesses in the construction sector through research*.

Jo, H., Han, I. and Lee, H. (1997) Bankruptcy prediction using case-based reasoning, neural networks, and discriminant analysis. *Expert Systems with Applications*. 13 (2), pp. 97–108.

Johnson, P. and Duberley, J. (2000) *Understanding Management Research: An Introduction to Epistemology*. London: SAGE Publications.

Johnson, R.B. and Onwuegbuzie, A.J. (2004) Mixed Methods Research: A Research Paradigm Whose Time Has Come. *Educational Researcher*. 33 (7), pp. 14–26.

Joy, O.M., Tollefson, J.O., Altman, E.I., Altman, E.I., Beaver, W.H., Dake, J.L., Edmister, R.O., Frank, R.E., Massy, W.F., Morrison, D.G., Johnson, C.G., Meehl, P.E., Rosen, A., Morrison, D.G., et al. (1975) On the Financial Applications of Discriminant Analysis. *The Journal of Financial and Quantitative Analysis*. 10 (5), pp. 723.

Kale, S. and Arditi, D. (1999) Age-dependent business failures in the US construction industry. *Construction Management and Economics*. 17 (4), pp. 493–503.

Kangari, B.R. and Farid, F. (1992) Financial performance analysis for construction industry. *Journal of Construction Engineering and Management*. 118 (2), pp. 349–361.

Kangari, R. (1988) Business failure in construction industry. *Journal of Construction Engineering and Management*. 114 (2), pp. 172–190.

Kapliński, O. (2008) *Usefulness and credibility of scoring methods in construction industry*. (January), pp. 37–41.

Keasey, K. and Watson, R. (1987) *Non-financial symptoms and the prediction of small company failure: a test of Argenti ' s hypotheses*. 14 (June 1986), .

Kim, S.Y. (2011) Prediction of hotel bankruptcy using support vector machine, artificial neural network, logistic regression, and multivariate discriminant analysis. *The Service*

Industries Journal. 31 (3), pp. 441–468.

Ko, L.-J., Blocher, E.J. and Lin, P.P. (2001a) Prediction of corporate financial distress: an application of the composite rule induction system. *The International Journal of Digital Accounting Research, ISSN 1577-8517, Vol. 1, Nº. 1, 2001, págs. 69-85*. 1 (1), pp. 69–85.

Ko, L.-J., Blocher, E.J. and Lin, P.P. (2001b) Prediction of Corporate Financial Distress: An Application of the Composite Rule Induction System. *The International Journal of Digital Accounting Research*. 1 (1), pp. 69–85.

Koksal, A. and Arditi, D. (2004) Predicting Construction Company Decline. *Journal of Construction Engineering and Management*. 130 (6), pp. 799–807.

Kołakowski, L. (1972) *Positivist Philosophy: From Hume to the Vienna Circle (Pelican)*. London: Penguin.

Korn, K.C. and Pine, J.B. (2014) *The Laws of Managing: Amazon.co.uk: Kim C Korn, Joesph B Pine II, B Joseph Pine II: 9781457532931: Books*. Indianapolis: Dog Ear Publishing.

Kuhn, T.S. (1962) *The Structure of Scientific Revolutions Second Edition, Enlarged The Structure of Scientific Revolutions*. Chicago: The University of Chicago Press.

Kumaraswamy, M.M. (1997) Conflicts, claims and disputes in construction. *Engineering, Construction and Architectural Management*. 4 (2), pp. 95–111.

Kuo, Y.-C. (2013) Consideration of Uneven Misclassification Cost and Group Size for Bankruptcy Prediction. *American Journal of Industrial and Business Management*. 3 pp. 708–714.

Van Laerhoven, H., van der Zaag-Loonen, H. and Derkx, B. (2004) A comparison of Likert scale and visual analogue scales as response options in children's questionnaires. *Acta Paediatrica*. 93 (6), pp. 830–835.

Laitinen, E.K. (1992) Prediction of failure of a newly founded firm. *Journal of Business Venturing*. 7 (4), pp. 323–340.

Langford, D., Iyagba, R. and Komba, D. (1993) Prediction of solvency in construction companies. *Construction Management and …*. 1993 (11), pp. 317–325.

Latham, G.P. and Finnegan, B.J. (1993) Personnel Selection and Assessment: Individual and Organizational Perspectives - Heinz Schuler, James L. Farr, J. Mike Smith - Google Books. In: Heinz Schuler, James L. Farr, and J. Mike Smith (eds.). *Personnel selection and assessment: Individual and organizational perspectives*. New Jersey: Lawrence Erlbaum Associates, Inc. pp. 41–45.

LaValle, S., Lesser, E. and Shockley, R. (2011) Big data, analytics and the path from insights to value. *MIT sloan*. 52 (2), pp. 21–32.

Lee, C.-W., Hsieh, K.-Y., Hsieh, S.-Y. and Hsiao, H.-C. (2014) A Dynamic Data Placement Strategy for Hadoop in Heterogeneous Environments. *Big Data Research*. 1 pp. 14–22.

Lei, H., Xing, T., Taylor, J. and Zhou, X. (2012) Monitoring Travel Time Reliability from the Cloud. *Transportation Research Record: Journal of the Transportation Research Board*. 2291 pp. 35–43.

Levinthal, D.A. and March, J.G. (1993) The myopia of learning. *Strategic Management Journal*. 14 (S2), pp. 95–112.

Lewin, A.Y., Long, C.P. and Carroll, T.N. (1999) The Coevolution of New Organizational Forms. *Organization Science*. 10 (5), pp. 535–550.

Li, B., Akintoye, A., Edwards, P.J. and Hardcastle, C. (2005) Critical success factors for PPP/PFI projects in the UK construction industry. *Construction Management and Economics*. 23 (5), pp. 459–471.

Liang, D., Tsai, C.-F. and Wu, H.-T. (2015) The effect of feature selection on financial distress prediction. *Knowledge-Based Systems*. 73 pp. 289–297.

Liaw, A. and Wiener, M. (2002) Classification and regression by randomForest. *R news*. 2 (3), pp. 18–22.

Madden, S. (2012) From Databases to Big Data. *IEEE Internet Computing*. 16 (3), .

Mahamid, I. (2012) Factors affecting contractor's business failure: contractors' perspective. *Engineering, Construction and Architectural Management*. 19 (3), pp. 269–285.

Mahout (2015) *Apache Mahout: Scalable machine learning and data mining*. Available from: https://mahout.apache.org/ [Accessed 5 November 2015].

Makeeva, E. and Neretina, E. (2013a) A Binary Model versus Discriminant Analysis Relating to Corporate Bankruptcies: The Case of Russian Construction Industry. *Journal of Accounting, Finance and Economics*. 3 (1), pp. 65–76.

Makeeva, E. and Neretina, E. (2013b) The Prediction of Bankruptcy in a Construction Industry of Russian Federation. *Journal of Modern Accounting and Auditing*. 9 (2), pp. 256–271.

Manning, C.D., Raghavan, P. and Schutze, H. (2008) *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

Marcella, R. and Illingworth, L. (2012) The impact of information behaviour on small business failure. *Information research*. 17 (9), .

March, J.G. (1981) Footnotes to Organizational Change. *Administrative Science Quarterly*. 26 (4), pp. 563.

Marshall, B., Cardon, P., Poddar, A. and Fontenot, R. (2013) Does Sample Size Matter in Qualitative Research?: A Review of Qualitative Interviews in is Research. *Journal of Computer Information Systems*. 54 (1), pp. 11–22.

Mason, R.J. and Harris, F.C. (1979) Predicting company failure in the construction industry. *Proceedings Institution of Civil Engineers*. 66 pp. 301–307.

Mays, N. and Pope, C. (1995) Rigour and qualitative research. *BMJ (Clinical research ed.)*. 311 (6997), pp. 109–112.

McGurr, P. and DeVaney, S. (1998) Predicting business failure of retail firms: an analysis using mixed industry models. *Journal of Business Research*. 43 (3), pp. 169–176.

Merton, R.K. and Kendall, P.L. (1946) The Focused Interview. *American Journal of Sociology*. 51 (6), pp. 541–557.

Miller, G.J. and Yang, K. (2007) *Handbook of Research Methods in Public Administration*. Florida: CRC Press.

Mintzberg, H., Ahlstrand, B. and Lampel, J. (1998) *Safari Strategy: A Guided Tour Through the Wilds of Strategic Management*. London: Prentice Hall.

Mitchell, J.C. (1983) Case and Situation Analysis. In: *Case Study Method*. London: SAGE Publications Ltd. pp. 165–186.

Mitkus, S. and Mitkus, T. (2014) Causes of Conflicts in a Construction Industry: A Communicational Approach. *Procedia - Social and Behavioral Sciences*. 110 pp. 777–786.

Mukherji, P. and Albon, D. (2010) *Research Methods in Early Childhood. An Introductory Guide*. London: SAGE.

Murphy, J. (1990) *Pragmatism: From Peirce to Davidson*. Colorado: Westview Press.

Muscettola, M. (2014) Probability of Default Estimation for Construction Firms - ProQuest. *International Business Research*. 7 (11), pp. 153–164.

Navon, R. (2005) Automated project performance control of construction projects. *Automation in Construction*. 14 (4), pp. 467–476.

Navon, R. (2007) Research in automated measurement of project performance indicators. *Automation in Construction*. 16 (2), pp. 176–188.

Ng, S.T., Wong, J.M.W. and Zhang, J. (2011) Applying Z-score model to distinguish insolvent construction companies in China. *Habitat International*. 35 (4), pp. 599–607.

Nicolás, J. and Toval, A. (2009) On the generation of requirements specifications from software engineering models: A systematic literature review. *Information and Software Technology*. 51 (9), pp. 1291–1307.

Nunnally, J.C. and Bernstein, I.H. (1994) *Psychometric Theory*. New York: McGraw-Hill.

Odom, M.D. and Sharda, R. (1990) A neural network model for bankruptcy prediction. In:

1990 IJCNN International Joint Conference on Neural Networks. 1990 San Diego, CA: IEEE. pp. 163–168 vol.2.

Odusami, K.., Iyagba, R.R.. and Omirin, M.. (2003) The relationship between project leadership, team composition and construction project performance in Nigeria. *International Journal of Project Management*. 21 (7), pp. 519–527.

Ohlhorst, F. (2013) *Big Data Analytics Turning Big Data into Big Money*. New Jersey: John Wiley and Sons.

Ohlson, J.A. (1980) Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research*. 18 (1), pp. 109.

Olavarrieta, S. and Ellinger, A.E. (1997) Resource-based theory and strategic logistics research. *International Journal of Physical Distribution and Logistics Management*. 27 (9/10), pp. 559–587.

Oyedele, L.O. (2013) Analysis of architects' demotivating factors in design firms. *International Journal of Project Management*. 31 (3), pp. 342–354.

Pallant, J. (2013) *SPSS Survival Manual*. Berkshire: McGraw-Hill Education.

Pearce, J.A. and Zahra, S.A. (1992) Board composition from a strategic contingency perspective. *Journal of Management Studies*. 29 (4), pp. 411–438.

Pflugfelder, E.H. and Helmut, E. (2013) Big data, big questions. *Communication Design Quarterly Review*. 1 (4), pp. 18–21.

Powell, T.C. (2001) Competitive advantage: logical and philosophical considerations. *Strategic Management Journal*. 22 (9), pp. 875–888.

Rae, D. (2000) Understanding entrepreneurial learning: a question of how? *International Journal of Entrepreneurial Behavior and Research*. 6 (3), pp. 145–159.

Ranger-Moore, J. (1997) Bigger may be better, but is older wiser? Organizational age and size in the New York life insurance industry. *American Sociological Review*. 62 (6), pp. 903–920.

Ravi Kumar, P. and Ravi, V. (2007) Bankruptcy prediction in banks and firms via statistical and intelligent techniques – A review. *European Journal of Operational Research*. 180 (1), pp. 1–28.

Razak Bin Ibrahim, A., Roy, M.H., Ahmed, Z.U. and Imtiaz, G. (2010) Analyzing the dynamics of the global construction industry: past, present and future. *Benchmarking: An International Journal*. 17 (2), pp. 232–252.

Reja, U., Manfreda, K.L., Hlebec, V. and Vehovar, V. (2003) Open-ended vs. Close-ended Questions in Web Questionnaires. *Developments in Applied Statistics Anuška Ferligoj and Andrej Mrvar (Editors) Metodološki zvezki*. 19 (1), pp. 160–117.

Remenyi, D., Williams, B., Money, A. and Shwartz, E. (1998) *Doing Research in Business and Management: An Introduction to Process and Method*. London: SAGE.

Resources, I. (2005) Research Methods Knowledge. *Practice*. pp. 199–201.

Rhodes, C. (2015) *Construction industry: statistics and policy.*

Richardson, F.M. and Davidson, L.F. (1984) On linear discrimination with accounting ratios. *Journal of Business Finance and Accounting*. 11 (4), pp. 511–525.

Ritzer, G. (2004) *Encyclopedia of Social Theory*. 2455 Teller Road, Thousand Oaks California 91320 United States: SAGE Publications, Inc.

Roberta, B. and Cowton, C.J. (2000) The E-Interview. *Forum: Qualitative Social Research*. 3 (2), .

Robinson, R.A. and Maguire, M.G. (2001) Top common causes of construction contractor failures. *Journal of Construction Accounting and Taxation*. Jan/Feb 20 (January), .

Robson, C.C.N.-N.I.B.R.O.B.I.B.R. (2011) *Real World Research : A Resource For Users of Social Research Methods In Applied Settings*. Oxford: Wiley.

Rolfe, G. (2006) Validity, trustworthiness and rigour: quality and the idea of qualitative research. *Journal of Advanced Nursing*. 53 (3), pp. 304–310.

Rooke, J., Seymour, D. and Crook, D. (1997) Preserving methodological consistency: a reply to Raftery, McGeorge and Walters. *Construction Management and Economics*. 15 (5), pp. 491–494.

Rosario, S.F. and Thangadurai, K. (2015) RELIEF: Feature Selection Approach. *International journal of innovative research and development*. 4 (11), .

Rosner, R.L. (2003) Earnings Manipulation in Failing Firms*. *Contemporary Accounting Research*. 20 (2), pp. 361–408.

Rumelt, R.P., Schendel, D. and Teece, D.J. (1991) Strategic management and economics. *Strategic Management Journal*. 12 (S2), pp. 5–29.

Russell, J. and Jaselskis, E. (1992) Predicting construction contractor failure prior to contract award. *Journal of Construction Engineering and Management*. 118 (4), pp. 791–811.

Russell, J.S. and Zhai, H. (1996) Predicting Contractor Failure Using Stochastic Dynamics of Economic and Financial Variables. *Journal of Construction Engineering and Management*. 122 (2), pp. 183–191.

Ryan, F., Coughlan, M. and Cronin, P. (2009) Interviewing in qualitative research: The one-to-one interview. *International Journal of Therapy and Rehabilitation*. 16 (6), pp. 309–314.

Sagiroglu, S. and Sinanc, D. (2013) Big data: A review. In: *2013 International Conference on Collaboration Technologies and Systems (CTS)*. May 2013 San Diego, CA: IEEE. pp. 42–47.

Saleem, Q., Ur, R. and Rehman, A. (2011) Impacts of liquidity ratios on profitability (Case of oil and gas companies of Pakistan). *Interdisciplinary Journal of Research in Business*. 1 (7), pp. 95–98.

Sánchez-Lasheras, F., de Andrés, J., Lorca, P. and de Cos Juez, F.J. (2012) A hybrid device for the solution of sampling bias problems in the forecasting of firms' bankruptcy. *Expert Systems with Applications*. 39 (8), pp. 7512–7523.

Saunders, M. and Paul, T. (2013) The Layers of Research Design | *Rapport* p.pp. 58–59.

Saunders, M.N.K., Lewis, P. and Thornhill, A. (2000) *Research Methods for Business Students*. Essex: Prentice Hall.

Schapire, R.E., Freund, Y., Bartlett, P., Lee, W.S. and Bar, P. (1998) Boosting the Margin: A New Explanation for the Effectiveness of Voting Methods. *Source: The Annals of Statistics The Annals of Statistics*. 26 (5), pp. 1651–1686.

Schlosser, R.W. (2007) Southwest Educational Development Laboratory. *Focus*. 17 (2007), pp. 1–8.

Scotland, J. (2012) Exploring the philosophical underpinnings of research: Relating ontology and epistemology to the methodology and methods of the scientific, interpretive, and. *English Language Teaching*. 5 (9), pp. 9–16.

Scott, J. (1981) *The probability of bankruptcy: A comparison of empirical predictions and theoretical models*. 5 pp. 317–344.

Sechrest, L. and Sidani, S. (1995) Quantitative and qualitative methods:: Is There an Alternative? *Evaluation and Program Planning*. 18 (1), pp. 77–87.

Senthamarai Kannan, S. and Ramaraj, N. (2010) A novel hybrid feature selection via Symmetrical Uncertainty ranking based local memetic search algorithm. *Knowledge-Based Systems*. 23 (6), pp. 580–585.

Seymour, D., Crook, D. and Rooke, J. (1997) The role of theory in construction management: a call for debate. *Construction Management and Economics*. 15 (1), pp. 117–119.

Seymour, D. and Rooke, J. (1995) The culture of the industry and the culture of research. *Construction Management and Economics*. 13 (6), pp. 511–523.

Siew, R.Y.J., Balatbat, M.C.A. and Carmichael, D.G. (2013) The relationship between sustainability practices and financial performance of construction companies. *Smart and Sustainable Built Environment*. 2 (1), pp. 6–27.

Singh, B., Kushwaha, N. and Vyas, O.P. (2014) A Feature Subset Selection Technique for High Dimensional Data Using Symmetric Uncertainty. *Journal of Data Analysis and Information Processing*. 2 (2), pp. 95–105.

Singh, D. and Tiong, R.L.K. (2006) Evaluating the financial health of construction contractors. *Proceedings of the Institution of Civil Engineers - Municipal Engineer*. 159 (3), pp. 161–166.

Smith, J.E. and Hakel, M.D. (1979) Convergence among data sources, response bias, and reliability and validity of a structured job analysis questionnaire. *Personnel Psychology*. 32 (4), pp. 677–692.

Smith, V., Devane, D., Begley, C.M., Clarke, M., Ghersi, D., Pang, T., Moher, D., Smith, V., Clarke, M., Smith, V., Williams, C., Becker, L., Oxman, A., Lichtenstein, A., et al. (2011) Methodology in conducting a systematic review of systematic reviews of healthcare interventions. *BMC Medical Research Methodology*. 11 (1), pp. 15.

Spector, P. (2015) *Summated Rating Scale Construction: An Introduction*. London: Sage Publications.

Stokes, D. and Blackburn, R. (2002) Learning the hard way: the lessons of owner-managers who have closed their businesses. *Journal of Small Business and Enterprise Development*. 9 (1), pp. 17–27.

Stroe, R. and Bărbuță-Mișu, N. (2010) Predicting the financial performance of the building sector enterprises — case study of galati county (Romania). *The Review of Finance and Banking*. 2 (1), pp. 29–39.

Sueyoshi, T. and Goto, M. (2009) DEA–DA for bankruptcy-based performance assessment: Misclassification analysis of Japanese construction industry. *European Journal of Operational Research*. 199 (2), pp. 576–594.

Sun, J., Liao, B. and Li, H. (2013) AdaBoost and Bagging Ensemble Approaches with Neural Network as Base Learner for Financial Distress Prediction of Chinese Construction and Real Estate Companies. *Recent Patents on Computer Science*. (1932), pp. 47–59.

Surety Information Office (2012) *Why Do Contractors Fail? Surety Bonds Provide Prevention and Protection National Association of Surety Bond Producers (NASBP) The Surety andamp; Fidelity Association of America (SFAA).*

Suthaharan, S. and Shan (2014) Big data classification. *ACM SIGMETRICS Performance*

Evaluation Review. 41 (4), pp. 70–73.

Tabachnik, B. and Fidell, L. (1989) *Using Multivariate Statistics*. Boston: Pearson/Allyn and Bacon.

Taffle, R.J. (1982) Forecasting Company Failure in the UK Using Discriminant Analysis and Financial Ratio Data. *Journal of Royal Statistical Society A*. 145 (3), pp. 342–358.

Taffler, R. (1983) The assessment of company solvency and performance using a statistical model.pdf. *Accounting and Business Research*. 15 (52), .

Tainton, B.E. (1990) The unit of analysis andquot;problemandquot; in educational research. *Queensland Journal of Educational Research*. 6 (1), pp. 4–19.

Talia, D. (2013) Toward cloud-based big-data analytics *IEEE Computer Science* p.pp. 98–101.

Tansey, P., Spillane, J.P. and Meng, X. (2014) Linking response strategies adopted by construction firms during the 2007 economic recession to Porter's generic strategies. *Construction Management and Economics*. 32 (7–8), pp. 705–724.

Tashakkori, A. and Teddlie, C. (1998) *Mixed Methodology: Combining Qualitative and Quantitative Approaches*. London: Sage.

Teece, D.J., Pisano, G. and Shuen, A. (1997) Dynamic Capabilities and Strategic Management. *Strategic Management Journal*. 18 (7), pp. 509–533.

The Construction Index (2011) *Bleak New Year for Failed Companies* -.

The Insolvency service (2016) *Insolvency Statistics – April to June 2016 (Q2 2016).*

Thompson, J.D. (2014) *Organizations in Action: Social Science Bases of Administrative Theory*. New Jersey: Transaction Publishers.

Torgo, L. (2011) *Data Mining with R: Learning with Case Studies*. Boca Raton, Florida: CRC Press.

Tranfield, D., Denyer, D. and Smart, P. (2003) Towards a Methodology for Developing Evidence-Informed Management Knowledge by Means of Systematic Review. *British*

Journal of Management. 14 (3), pp. 207–222.

Trochim, W.M.K. (2006) *Research Methods Knowledge Base*. Ohio: Atomic Dog/Cengage Learning.

Tsai, L.-K., Tserng, H.-P., Liao, H.-H., Chen, P.-C. and Wang, W.-P. (2012) Integration of Accounting-Based and Option-Based Models to Predict Construction Contractor Default. *Journal of Marine Science and Technology*. 20 (5), pp. 479–484.

Tseng, F.-M. and Hu, Y.-C. (2010) Comparing four bankruptcy prediction models: Logit, quadratic interval logit, neural and fuzzy neural networks. *Expert Systems with Applications*. 37 (3), pp. 1846–1853.

Tserng, H.P., Chen, P.-C., Huang, W.-H., Lei, M.C. and Tran, Q.H. (2014) Prediction of default probability for construction firms using the logit model. *Journal of Civil Engineering and Management*. 20 (2), pp. 247–255.

Tserng, H.P., Liao, H.-H., Jaselskis, E.J., Tsai, L.K. and Chen, P.-C. (2012) Predicting Construction Contractor Default with Barrier Option Model. *Journal of Construction Engineering and Management*. 138 (5), pp. 621–630.

Tserng, H.P., Liao, H.-H., Tsai, L.K. and Chen, P.-C. (2011a) Predicting Construction Contractor Default with Option-Based Credit Models—Models' Performance and Comparison with Financial Ratio Models. *Journal of Construction Engineering and Management*. 137 (6), pp. 412–420.

Tserng, H.P., Lin, G.F., Tsai, L.K. and Chen, P.C. (2011b) An enforced support vector machine model for construction contractor default prediction. *Automation in Construction*. 20 (8), pp. 1242–1249.

Tserng, H.P., Ngo, T.L., Chen, P.C. and Quyen Tran, L. (2015) A Grey System Theory-Based Default Prediction Model for Construction Firms. *Computer-Aided Civil and Infrastructure Engineering*. 30 (2), pp. 120–134.

Tucker, J. (1996) Neural networks versus logistic regression in financial modelling: a methodological comparison. In: *Proceedings of the 1996 World First Online Workshop on Soft Computing*. 1996

Vapnik, V.N. (1998) *Statistical learning theory*. San Francisco: John Wiley and Sons.

Wang, G., Ma, J. and Yang, S. (2014) An improved boosting based on feature selection for corporate bankruptcy prediction. *Expert Systems with Applications*. 41 (5), pp. 2353–2361.

Watson, J. and Everett, J. (1993) Defining Small Business Failure. *International Small Business Journal*. 11 (3), pp. 35–48.

Wedzki, D. (2005) Multivariate analysis of bankruptcy on the example of building industry. *Operations Research and Decisions*. 2 pp. 59–81.

Wei, S. and Lin, Z. (2010) *Accelerating Iterations Involving Eigenvalue or Singular Value Decomposition by Block Lanczos with Warm Start*.

Wernerfelt, B. (1984) A resource-based view of the firm. *Strategic Management Journal*. 5 (2), pp. 171–180.

Wholey, D.R. and Brittain, J.W. (1986) Organizational Ecology: Findings and Implications. *Academy of Management Review*. 11 (3), pp. 513–533.

Wright, K.B. (2005) Researching Internet-Based Populations: Advantages and Disadvantages of Online Survey Research, Online Questionnaire Authoring Software Packages, and Web Survey Services. *Journal of Computer-Mediated Communication*. 10 (3), pp. 00–00.

Xindong Wu, Xingquan Zhu, Gong-Qing Wu and Wei Ding (2014) Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*. 26 (1), pp. 97–107.

Yeh, C.-C., Chi, D.-J. and Lin, Y.-R. (2014) Going-concern prediction using hybrid random forests and rough set approach. *Information Sciences*. 254 pp. 98–110.

Yin, R.K. (2003) *Methods, Case Study Research: Design and Research*. London: Sage Publications.

Yoon, J.S. and Kwon, Y.S. (2010) A practical approach to bankruptcy prediction for small businesses: Substituting the unavailable financial data for credit card sales information. *Expert Systems with Applications*. 37 (5), pp. 3624–3629.

Young, B. (2001) Postpositivism and educational research. *The International Journal of Educational Management*. 15 (March), pp. 240–240.

Zaharia, M., Chowdhury, M., Franklin, M. and Shenker, S. (2010) Spark: cluster computing with working sets. In: *HotCloud*. 2010

Zavgren, C. (1983) The prediction of corporate failure: the state of the art. *Journal of Accounting Literature,*. 2 (1), pp. 1–38.

Zavgren, C. V. (1985) Assessing the Vulnerability to Failure of American Industrial Firms: A Logistic Analysis. *Journal of Business Finance and Accounting*. 12 (1), pp. 19–45.

Zhanquan, S. and Fox, G. (2012) *Large Scale Classification Based on Combination of Parallel SVM and Interpolative MDS*.

Zhao, Z., Zhao, X., Davidson, K. and Zuo, J. (2012) A corporate social responsibility indicator system for construction enterprises. *Journal of Cleaner Production*. 29–30 pp. 277–289.

Zhou, L., Lai, K.K. and Yen, J. (2014) Bankruptcy prediction using SVM models with a new approach to combine features selection and parameter optimisation. *International Journal of Systems Science*. 45 (3), pp. 241–253.

Zikopoulos, P. and Eaton, C. (2011) *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. New York: McGraw-Hill Osborne Media.

APPENDIX A

A1: Example questionnaire addressed to a top management team member of a failed construction firm

Address

Dear (name of identified respondent)

Factors Affecting Failure/Insolvency of UK Construction Firms

I am a doctoral researcher at the Bristol Enterprise Research and Innovation Centre (BERIC) in University of the West of England (UWE) under the supervision of Professor Lukumon Oyedele. I am researching the factors that contribute to the failure/insolvency of construction firms and this questionnaire is created to provide the necessary information to complete my research successfully. This questionnaire specifically requires responses from especially owners, directors or management level staff of existing and/or insolvent (or dormant, failed, etc.) construction firms. I believe you have the experience required to complete this questionnaire because you have been identified as a former director of the now defunct (name of failed construction firm) and a current director of the existing (name of existing construction firm where respondent is currently a TMT member) through a financial database. I am thus using this opportunity to plea with you to please help me complete this questionnaire in relation to (name of failed construction firm). I assure and guarantee you that all information provided will be kept confidential. A free return envelope with my supervisor's (Professor Lukumon Oyedele) address as correspondent address has been enclosed with this questionnaire. Thank you very much for your anticipated contribution.

Hafiz Alaka
Doctoral Researcher | Bristol Enterprise, Research and Innovation Centre (BERIC)
University of the West of England, Bristol
Email: Hafiz2.Alaka@live.uwe.ac.uk
Tel: +44(0)7574819428 | +44 (0)7535018889

Professor Lukumon O. Oyedele
Director of Bristol Enterprise, Research and Innovation Centre (BERIC)
Bristol Business School
University of the West of England, Bristol
Frenchay Campus
Bristol BS16 1QY
E-mail: L.Oyedele@uwe.ac.uk

Tel: +44 (0) 117 32 83443

<u>Section A – Respondent's Details</u>

Please mark answers with an 'x' where tick boxes are provided

1. Name of construction firm your answers are based on **(confidentiality is assured)**. (name of failed construction firm).
2. Number of employees that works/worked for firm: □ 1-10 □ 11-50

   □ 51-250 □ over 250
3. Total number of years of construction industry experience of respondent

   □ 1-5 □ 6-10 □ 11-15 □ 16-20 □ 21-25 □ 26-30 □ 31-35 □ 36-40 □ over 40
4. Position(s) respondent holds/held in the firm (tick as many as applicable):

   □ Owner □ CEO/MD/President/CE □ Chairman □ Director

   □ Board member □ Senior manager □ Project manager

   □ others (please specify) ____________________
5. Highest Qualification of respondent:

   □ A-Level □ HND □ Degree □ Masters □ PhD
   □ others (please specify) ___________
6. How many branch offices does/did the firm have? _________________________

<u>Section B – Top Management Characteristics</u>

Please note: CEO = Chief executive officer/President/MD/Chief executive (or owner, where the owner is the CEO) of the firm

1. Age of CEO? □ 16-20 □ 21-30 □ 31-40 □ 41-50 □ 51-70

   □ above 70
2. Gender of CEO □ Male □ Female □ Others
3. Nationality of CEO (if not sure, then fill in the continent CEO):

   ____
4. Highest Qualification of CEO

   □ A-Level □ HND □ Degree □ Masters □ PhD

   □ others (please specify) ___________
5. Does CEO have any form of management training? □ Yes □ No
6. Number of years CEO has spent with firm

   □ 1-5 □ 6-10 □ 11-15 □ 16-20 □ 21-25 □ 26-30

   □ 31-35 □ 36-40 □ over 40
7. Total number of years of construction industry experience of CEO

□ 1-5 □ 6-10 □ 11-15 □ 16-20 □ 21-25 □ 26-30
□ 31-35 □ 36-40 □ over 40

8. Profession of CEO
□ Builder □ Civil/Structural engineer □ Architect
□ Quantity surveyor □ Land surveyor □ Project manager
□ Accountant □ Others (please specify) ____________________

Section C – Senior Management and Finance Questions

	Please answer Y = Yes or N = No for the following questions	**Y**	**N**
C1.	The firm is/was owned by a single person		
C2.	The owner is/was the same person as the chief executive (CEO)/president/Managing Director (MD) of the firm		
C3.	The firm has/had a board of directors		
C4.	If yes, how many directors does/did the firm have?		
C5.	The firm took over of another firm at some point in time		
C6.	If yes, was the take over as a result of financial or other types of distress?		
C7.	The firm has/had a clear bidding strategy		
C8.	There is/was a clear sub-contractor selection process		
C9.	The firm has/had a long term strategic goal		
C10.	The firm is/was specialized in a particular trade or service		
C11.	Has the range of trade/services broadened over time		
C12.	The firm change its main specialization of construction work (e.g. from public to private project, or from building residential homes to commercial stores, etc.) at some point in time		
C13.	The owner is/was on a fixed salary		
C14.	There is/was a dedicated financial director		
C15.	The financial director is/was performing another role at the same time		
C16.	The company account is/was clearly separated from any personal accounts		
C17.	Was account management fully computerized		
C18.	The firm consistently run/ran negative cash flow		
C19.	The firm went through an expansion programme less than 2 years ago or within 2 years before closing down		

<u>Section D</u> – <u>Proportion of firms' professionals with high qualifications/skills and involvement</u>

Please indicate the percentage proportion/fraction of the following factors in your firm, giving actual or approximate answers. The scale of relevance is 1-5 where:

1 = 0-20% 2 = 21-40% 3 = 41-60% 4 = 61-80% 5 = 81-100%

	What fraction of the following factors exists/existed in the firm	**Factor Fraction**				
		1	**2**	**3**	**4**	**5**
D1.	Percentage of passive members in the board of directors					
D2.	Percentage of directors that worked in the firm					
D3.	Percentage of directors that had construction background					
D4.	Percentage of directors that had management/administrative background					
D5.	Percentage of directors educated to at least a degree level					
D6.	Percentage of personnel educated to at least a degree level					
D7.	Percentage of works usually subcontracted during projects					
D8.	Percentage of successful bids					
D9.	Percentage of firm's earnings invested in properties					
D10.	Percentage of firm's earnings used in construction operations					
D11.	Percentage of professional workers that were registered with professional bodies					

<u>Section E –</u> <u>The effect of external, industrial and firm characteristic factors</u>

Please ignore questions that do not apply to the firm. Please indicate how the following factors have affected the firm, giving actual or approximate answers. The scale of relevance is 1-5 where:

1 = Very negatively 2 = Negatively 3 = No real effect 4 = Positively 5 = Very Positively

	How are/have any of the following factors affecting/affected your firm?	**Effect of Factor**				
		1	**2**	**3**	**4**	**5**
E1.	The 2008 global financial crises [Economic recession(s)]					
E2.	High immigration levels in UK					
E3.	Influx of firms into the industry, (from across the country and outside the country)					
E4.	Fluctuation in construction material costs					
E5.	Construction industry culture					
E6.	Construction industry environmental sustainability agenda					

E7.	Type/Quality of workforce available for employment					
E8.	Newness [i.e. how did newness (first four years) affect the performance of the firm in its early years?]					
E9.	The company size					
E10.	Fraud (if fraud ever happened, how it affected the firm?)					
E11.	Natural disasters (whether directly on the firm or its projects)					

Section F – Frequency of occurrence of some project related factors

Please ignore questions that do not apply to the firm. Please mark the frequency of occurrence of the following factors in your firm. The scale of relevance is 1-5 where:

1 = Not at all 2 = Rarely 3 = Sometimes 4 = Fairly often 5 = Very often

	How often do/did the following factors happen in the firm?	**Factor Frequency**				
		1	**2**	**3**	**4**	**5**
F1.	Very late collection of payment for completed works					
F2.	Unsuccessful collection of payment for completed works					
F3.	Get cash strapped on projects (cash flow)					
F4.	Reach debt limit with bank/financier					
F5.	Renegotiate loan terms					
F6.	Make profit on projects					
F7.	Produce complete financial statements					
F8.	Bid for jobs outside firm's specialty					
F9.	Executed project cost more than the bidding price used to win contract					
F10.	Submit very low bids because of fierce competition					
F11.	Rely on government projects					
F12.	Rely on private projects					
F13.	Firm win major bids it submitted					
F14.	Firm completes project within stipulated time frame					
F15.	Firm completes project within bidding budget					
F16.	Firm executes project to time and cost without conflict					
F17.	Internal conflict arises within the firm					
F18.	Internal conflict within the organization gets uncomplicatedly resolved					
F19.	Firm gets project through referral from another customer					
F20.	Expansion of firm					

	How often do/did the following factors happen in the firm?	**Factor Frequency**				
		1	**2**	**3**	**4**	**5**
F21.	Conflicts with clients on projects					
F22.	Conflicts with subcontractor in terms of subcontractors not showing up, performing low quality works.					
F23.	Delay of payments to subcontractors.					
F24.	Conflicts with other major parties on projects					
F25.	Conflict /litigation/legal issues / dispute arise from completed projects					
F26.	Losing out in conflict /litigation/legal issues /dispute cases					
F27.	Customers offer repeat business					
F28.	Repeated use of particular sub-contractor(s)					
F29.	Materials are supplied to firm on credit					
F30.	Debts payment to suppliers are delayed					
F31.	Legal advice sorted for contracts taken					
F32.	Problems with labour cost					
F33.	Execution of multiple projects simultaneously					
F34.	Bid for projects outside main geographical area of comfort (city, county, region, etc.)					
F35.	Register accidents on its site					
F36.	Replace key personnel					
F37.	Execute a highly financially challenging project					

Section G – The characteristics and performance level of the firm, its management and its staff

Please note: CEO = Chief executive officer/President/MD/Chief executive (or owner, where the owner is the CEO) of the firm

Please ignore questions that do not apply to the firm. Please mark the extent to which the firm and/or is staff exhibit(ed)/perform(ed) each of the following factors, giving actual or approximate answers. The scale of relevance is 1-5 where:

1 = Very low 2 = Low 3 = Moderate 4 = High 5 = Very high

	Please rate the performance of the firm and its staff with regards to the following factors?	**Performance Level**				
		1	**2**	**3**	**4**	**5**
G1.	Enthusiasm of the project management team					
G2.	Level of overall competence of top management team					

	Please rate the performance of the firm and its staff with regards to the following factors?	**Performance Level**				
		1	**2**	**3**	**4**	**5**
G3.	The willingness of the top management team to take risk					
G4.	The motivation of the CEO/directors					
G5.	The tolerance of the CEO					
G6.	The decisiveness of the CEO/directors					
G7.	Leadership support of CEO/directors to employees					
G8.	The creativity/innovation of the CEO/directors					
G9.	The integrity/transparency of the CEO/directors					
G10.	The flexibility of the CEO/directors					
G11.	The reliability/dependability of the CEO/directors					
G12.	The construction industry knowledge of the CEO/directors of the firm					
G13.	The CEO’s/directors’ ‘response to feedback’					
G14.	Commitment of project management team					
G15.	Level of firm’s response to market change					
G16.	The effectiveness of the financial director					
G17.	The profit levels of the firm					
G18.	The liquidity level of the firm					
G19.	Firm’s reception to latest technologies					

Please write any additional comments in the box below (You can staple additional sheet if need be)