Applied Computing and Informatics xxx (2018) xxx–xxx

Contents lists available at ScienceDirect

Applied Computing and Informatics

journal homepage: www.sciencedirect.com

Original Article

A framework for big data analytics approach to failure prediction of construction firms

Hafiz A. Alaka [a], Lukumon O. Oyedele [b,*], Hakeem A. Owolabi [c], Muhammad Bilal [d], Saheed O. Ajayi [e], Olugbenga O. Akinade [d]

[a] Coventry University, United Kingdom
[b] Bristol Enterprise Research and Innovation Centre, University of the West of England, Bristol, United Kingdom
[c] The University of Northampton, United Kingdom
[d] University of the West of England, Bristol, United Kingdom
[e] Leeds Beckett University, United Kingdom

ARTICLE INFO

Article history:
Received 14 September 2017
Revised 9 April 2018
Accepted 11 April 2018
Available online xxxx

Keywords:
Big data analytics
Failure prediction models
Construction businesses
Machine learning
MapReduce/Spark

ABSTRACT

This study explored use of big data analytics (BDA) to analyse data of a large number of construction firms to develop a construction business failure prediction model (CB-FPM). Careful analysis of literature revealed financial ratios as the best form of variable for this problem. Because of MapReduce's unsuitability for iteration problems involved in developing CB-FPMs, various BDA initiatives for iteration problems were identified. A BDA framework for developing CB-FPM was proposed. It was validated by using 150,000 datacells of 30,000 construction firms, artificial neural network, Amazon Elastic Compute Cloud, Apache Spark and the R software. The BDA CB-FPM was developed in eight seconds while the same process without BDA was aborted after nine hours without success. This shows the issue of not wanting to use large dataset to develop CB-FPM due to tedious duration is resolvable by applying BDA technique. The BDA CB-FPM largely outperformed an ordinary CB-FPM developed with a dataset of 200 construction firms, proving that use of larger sample size with the aid of BDA, leads to better performing CB-FPMs. The high financial and social cost associated with misclassifications (i.e. model error) thus makes adoption of BDA CB-FPMs very important for, among others, financiers, clients and policy makers.

1. Introduction

The construction industry remains a major player of any country's economy. The significance of any country's (or region's) economy is such that the sustainable development of the country largely hinges upon it [1]. A country's economy, alongside its military might (which is also partly based on economic power), have been identified as the two key necessities that must be at a relatively superior level for the country to be considered a 'superpower'. A poor economy on the other hand leads to poverty which in turn leads to lack of basic amenities, high crime rate, low life expectancy, etc. [2]. In a nutshell, the absolute significance of a country's economy cannot be overemphasized hence anything that contributes to, or affects, it significantly is usually of national/-global concern.

The United Kingdom's Department for Business Innovation and Skills [3] clearly stated that the construction sector is among the biggest sectors of the UK economy. The department 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 [4] in a House of

* Corresponding author.
E-mail addresses: ac7485@coventry.ac.uk (H.A. Alaka), L.Oyedele@uwe.ac.uk (L.O. Oyedele), hakeem.owolabi@northampton.ac.uk (H.A. Owolabi), muhammad.bilal@uwe.ac.uk (M. Bilal), S.Ajayi@leedsbeckett.ac.uk (S.O. Ajayi), olugbenga.akinade@uwe.ac.uk (O.O. Akinade).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.aci.2018.04.003
2210-8327/

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.

In the European union (EU), the CI boasts 20 million direct employees representing 15% of all EU employees; this equates to over 10% of the EU GDP and over 50% of its fixed capital formation [5]. It (i.e. the CI) represents the biggest lone economic activity and affects 44 million employees directly or indirectly [5]. 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 [6] (and $15.5 trillion by 2030 [7]).

The industry has however consistently led, or been around the top of the insolvency chart by sector in most countries, including the United Kingdom [8,9], thereby causing serious troubles for many economies. In 2012 the construction sector insolvency rate in the UK was third highest at 14.4 percent [9]. In England and Wales alone, construction businesses made up 23% of the total number of all businesses forced into compulsory liquidation in 2012 [10]. More 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 [8]. This included 833 obligatory or forced liquidations and 2143 unforced liquidations. According to European Commission (2012), many construction companies in Europe are folding up, significantly downsizing or shifting attention to other parts of construction they did not used to deal with. In spite of a relatively bettering global economy in recent times, the CI still had the highest percentage of failed business in the world at 20.2 percent in 2012 according to Dun and Bradstreet [9].

One main step commonly taken to reduce failure of construction firms is the development of construction businesses failure prediction models (CB-FPM). The reliable performance of these CB-FPMs is however partly dependent on the size of data used to develop them. To develop a highly reliable CB-FPM, a relatively large dataset containing data of tens of thousands of construction firms might be needed. Some studies have attempted using a relatively large dataset but fall very short of tens of thousands of firms. Van Frederikslust [12], for example, used data of 40 (20 failed and 20 surviving) sample firms over a 20 year period equating to a set of nearly 800 yearly financial reports. Altman et al. [13] went much further by using the data of a thousand firms to develop their failure prediction models (FPM). More recently, Chen [14] was able to generate 1615 financial statements from 42 sample construction firms. Though these datasets can be considered as being relatively large, they are far too small to develop a very reliable CB-FPM.

The authors of the cited studies might have well been conscious and cautious of the fact that the available tools will go through a long tedious computational duration when analysing really large data. An example of this long duration is Odom & Sharda [15] model development which took 24 h to build using ANN. Altman et al. [13] also reported significant machine hours for ANN training on a thousand firms data. Overcoming this long, tedious duration problem is the main motivation of this study. It is believed that this can be done by using the novel big data analytics (BDA) technology. However, using BDA to develop CB-FPM is not a straight forward process because of the iterative process required in classification analytics, which is what is employed to develop CB-FPM. The objectives of this study are thus:

- To propose a framework architecture for the development process of a big data analytics (BDA) CB-FPM.
- To implement the framework in developing a BDA CB-FPM

This main contribution of this study is the development and validation of a framework architecture for using big data analytics (BDA) to develop a CB-FPM. The framework will help to suppress the unappealing computational intensity of using large data set to develop CB-FPM. It will effectively eliminate the long tedious computational duration normally associated with using network algorithms like ANN and Bart (Bayesian Additive Regression Trees) machine with large data.

The next section is a literature review of CB-FPM studies. Section 3 is an explanation of what 'big data' is and the fitness of CB-FPM data for big data analytics. Section 4 discusses the suitable variable type, potential sources of data and data challenges for a CB-FPM to be developed with the big data technology. Section 5 details the BDA initiatives that might be suitable for developing CB-FPM. Section 6 proposes a framework architecture for developing CB-FPM while Section 7 contain details of implementation of part of the framework to develop a CB-FPM. Section 8 draws the conclusion to the study.

2. Literature review

The study of failure prediction of companies dates back to 1966 when Beaver [16], in a novel study, used a univariate system of financial ratios to attempt prediction of bankruptcy of firms. "A financial ratio is a quotient of two numbers, where both numbers consist of financial statement items" [16,p. 71–72]. This study was followed by Altman's [17] multivariate approach. Altman employed the multi-discriminant analysis (MDA) statistical tool and a set of financial variables for his prediction and most failure prediction studies since then have adopted this approach. However, some of the succeeding studies, especially the most recent ones since around 2006, have used machine learning tools like ANN.

The first study to develop a failure prediction model for construction firms was authored by Fadel [18] who used profitability ratios as variables with the MDA tool in a pilot study. Mason and Harris [19] later developed a proper model using MDA and six financial variables. The aim of the project, according to the authors, was to check Altman's technique on predicting failure of construction firms. Using the data of just 40 construction firms, Mason and Harris [19] achieved an overall accuracy of 87% and concluded the technique was valid. Kangari and Farid [20] used a slightly different approach by combining six financial variables with non-financial variables like company size. As opposed to using MDA, they used the logistic regression (LR) statistical technique. They however did not state the sample size used. Langford, Iyagba and Komba (1993) tested Altman's model on three construction firms, got an accuracy of 63.3% and concluded that the construction industry needed its own specific models.

Later studies increased sample sizes to improve reliability. For examples Abidali and Harris [22] and Russell and Zhai [23] used the data of 31 and 120 construction firms and achieved accuracy values 70.3% and 78.3% respectively. Both studies used the MDA technique. However, Alaka et al. [24] noted that the MDA and LR techniques have many assumptions which the mentioned studies did not satisfy before using them.

At the turn of the century, various other techniques were trialled for developing CB-FPMs. Singh and Tiong [25] trialled the entropy technique, while Huang [26] trialled the Structural models of credit risk but none of these techniques gained wide acceptance. Sueyoshi and Goto [27] also tried a different variant of the MDA which they labelled Data Envelopment Analysis–Discriminant Analysis (DEA–DA). Of these three studies, Sueyoshi and Goto [27] used the largest sample size, which consisted of 215 sample Japanese construction firms.

Although the first study to use machine learning tools for an FPM was in 1990 by Odom and Sharda [15], CB-FPM studies did not start using them until Tserng et al. [28] used an enforced

support vector machine to develop FPMs for construction contractors on the New York Stock Exchange (NYSE), American Exchange (AMEX), and Nasdaq. Tserng et al. [28] utilized only a total of 168 construction contractors as their sample. Although many recent CB-FPM studies (after 2010) still used the statistical (MDA and LR) techniques [e.g. 29–32], a few others have adopted machine learning tools [14,e.g. 33–35]. Of all these studies, Heo and Yang [35] used the largest sample size which consisted of 2762 construction firms. This is despite the fact that a larger sample size helps to improve reliability [36].

The associated long duration computation cost which comes with using a really large sample with ML tools, as very evident in Du Jardin's [37] study, makes it understandable that CB-FPM researchers avoid it. Contemporary technology like big data analytics, which are built to deal with large data, should be able to reasonably reduce this long duration. This study thus sets out to use a relatively large sample size, and use BDA to avoid the potentially associated tedious duration. However, BDA executed with the popular MapReduce framework is not built for iterative process required during CB-FPM development, hence BDA application in this field is not straight forward. This study hence sets out to create a framework architecture for the development process of a big data analytics (BDA) CB-FPM, and to test the framework by implementing it.

3. Big data analytics and the suitability of CB-FPM data

Big data has generally been defined in relation to three main feature: volume, variety and velocity [38]. Volume deals with size usually, but not always, in terabytes or petabytes of data and beyond. Velocity has to do with the speed with which data is generated while variety refers to the variability in the format of data (e.g. picture file, text file, audio file, etc.). Apache Hadoop is probably the most popular and complete big data framework presently.

Apache Hadoop, which has four major components as explained below, can be described as a comprehensive free of charge big data setup for distributed and scalable computing (Fig. 1).

1. Hadoop common: this consist of the utilities and libraries needed by other Hadoop components
2. HDFS: a file system with numerous nodes on which huge data can be deposited so that analysis can take place simultaneously on different nodes as if they are on a single computer [39].
3. Hadoop Yarn: a structure that takes care of how data is distributed to nodes during analysis [40]
4. MapReduce: this "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" [41], p.107] (see Fig. 2).

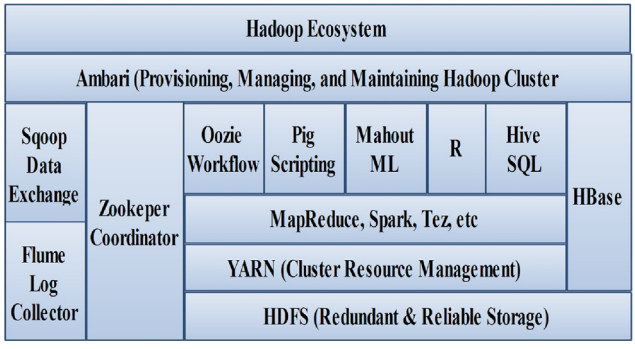

Fig. 1. The Hadoop network of interconnected system.

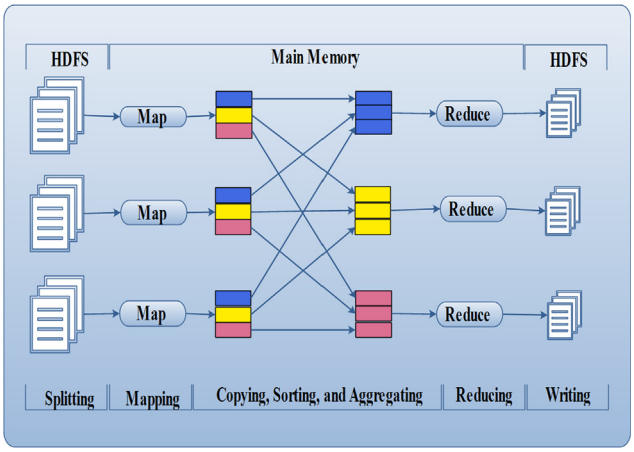

Fig. 2. How the MapReduce function works.

There have been various advances in the field of big data analytics research. Some of the recent ones include improving security of the data to be analysed with big data analytics system set up on cloud [42,43]. This is because cloud owners like Amazon or Microsoft, for example, have access to data stored in their cloud. As MapReduce is not capable of handling iteration problems very well and is thus unfit for developing CB-FPMs, some advances in terms of BDA initiatives fit for iteration problems have also been made (please see Section 5 for details of these advances).

3.1. Fitness of CB-FPM data to big data analytics

While big data used to be defined mainly around three features (volume, variety and velocity), as initially highlighted, a new feature that has been added to that is the nature of analysis [44,45]. This is why big data is defined in some avenues as data that cannot be analysed using conventional computer systems. Although some data are relatively big, their size do not qualify to be big data in terms of volume and they can thus be processed on a normal computer. However, certain types of analyses that require tedious computations might not be performable on such data on normal computers. The best example of this is Jacob's [44] experiment where he was able to use a normal computer to perform simple analysis on a created demographic data of world population in a table of circa 10 columns and over 7 billion rows which was contained in a 100 gigabyte hard disk. He was unable to load the same data unto enterprise grade database system (PostgreSQL6) using a super performance computer even before any analysis. This data is not qualified as big data for the first analysis but is qualified as big data in the case of the second unsuccessful attempted analysis.

The above example is what makes the data of tens of thousands of construction firms 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 with a machine learning tool like artificial neural network (ANN) which usually performs a tediously large number of iterations to achieve convergence will potentially take numerous hours on a normal computer as with Du Jardin's (2010) study. Such analysis hence qualifies the data for big data analytics.

Please cite this article in press as: H.A. Alaka et al., Applied Computing and Informatics (2018), https://doi.org/10.1016/j.aci.2018.04.003

4. Variables and potential data sources for big data analytics CB-FPM, and challenges

4.1. Financial ratios of construction firms as variables

The vast majority of FPMs developed for both construction and non-construction businesses have used only financial ratios as variables [e.g. 13,21,22,46–48, among many others]. The extensive use of financial ratios is because some of the pioneering works used only financial ratios, the pioneers were account/finance professionals [16,49] and most importantly, financial ratios are readily available from periodically published financial statements of firm's making access to such data relatively easy for a FPM developer.

However, countless number of non-financial indications of construction firms insolvency, such as management mistakes, do come up a lot earlier than the financial distress shown by financial ratios [22]. Financial distress only tends to be noticeable when the failure process is almost complete, around the last two years of failure. In fact, it is adverse managerial actions and other qualitative factors that normally lead to poor financial standing and in turn cause insolvency. It follows that managerial decisions, company activities, etc. (qualitative variables) influence the results of financial ratios hence for early prediction, which is the aim of many prediction models in order to allow enough time for remedy, the employment of qualitative variables is necessary [20,50–53, among others]. They are however neither readily decided nor readily available. Further, data collection for non-financial variables usually involves interviewing respondents or sending out questionnaires. This can be really difficult where the number of respondents required is in tens of thousands. Non-financial variables will thus be hard to get in large volumes for the purpose of developing CF-BPM using big data analytics.

Conclusively, for big data analytics CB-FPM, only construction firms' financial ratios as quantitative variables will be viable [54].

4.2. Potential data sources

The sources for financial information (or financial ratios) for public companies are quite simple to identify. A data source like DataStream hosts the financial information of many public construction companies around the world (including US and Europe). For country-specific information, FAME (Financial Analysis Made Easy) Bureau Van Dijk database offers financial information of over 600,000 construction firms in UK for example. Financial information of over 3 million companies in the UK can also be gotten from Company House. Over 14 years of balance sheet and financial reports of many US construction companies can be downloaded from Compustat, Mergent Online and Mergent/Moody's Online Manuals. A simple Google search using the search words 'financial databases list' will return many pages with plenty of such data sources. Virtually all these data sources can export data direct to excel hence exporting the required data should not be a very big problem. DataStream even has an excel add-in that allows some direct searches of its database and direct analysis through excel.

4.3. Data challenges

The first main challenge is how to go about downloading or exporting data of tens of thousands of construction firms to excel one after the other. This will take too long or require the services of so many people to actualise. The same problem applies to merging the data of the tens of thousands of firms into one or more excel sheets in a structured way, if necessary, before uploading to the platform where the big data analysis will take place. One solution to this challenge is for data sources to allow a direct download of the financial ratios of all the companies returned in a particular search into one Excel sheet in a structured way. For example, a search for all the construction companies that have failed (receivership, dormant, dissolution, liquidation and inactive) in the UK since the start of the years 2001 and 2015 on FAME yielded a result of nearly 260,000 companies (see Fig. 3). Having a command that will allow the data of all these companies to be downloaded into their separate files at once, or into a single Excel sheet in a structured manner, will solve this challenge. Another option is to use the SQL language to query the data sources.

Another set of challenges are the potential uncertainty and incompleteness of information from data sources. For example, some financial ratios can be missing from some reports, the report of some construction firms might be missing details of a year or more, etc. Also, the data might not readily differentiate between data for failed and existing firms as is normally needed in supervised learning which is more commonly used for construction CB-FPM. To overcome this challenge, it could be decided that only firms with complete data will be used in developing a CB-FPM. This is however difficult in the case of data of tens of thousands of construction firms because the total number of construction firms with complete financial data, as observed from data sources, is barely up to five thousand. An easier way of solving this problem is to employ techniques that can be used to produce values for missing data.

5. Machine learning tools for big data analytics CB-FPM

The machine learning (ML) tools used for developing CB-FPMs will struggle when it comes to carrying out a robust analysis on any huge data that might require more than a single machine's memory for analysis [55] hence it is almost impractical to use using ML tools for a direct analysis huge data [56]. Further, ML tools are not very compatible with MapReduce, especially when the computation in question involves iteration [57]. Consequently, various BDA initiatives which support iteration have been developed.

Many of the BDA initiatives that support iteration are MapReduce based apart from Apache Mahout Spark model. Some MapReduce related initiatives include Indiana University's Twister, University of Washington's HaLoop and Microsoft's Project Daytona. These initiatives are available for use free of charge.

BDA initiative selection for CB-FPM development is dependent on certain features including location of data, the distributed file

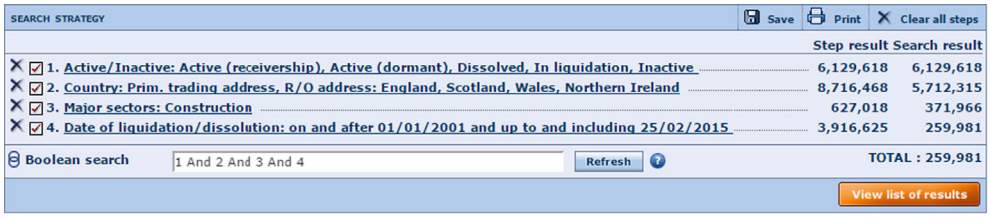

Fig. 3. An example search result in FAME yields a large number of construction firms.

system to be used, among others. Except for Apache Spark, all initiatives have this restriction. For instance, data needs to be uploaded to Azure cloud for Daytona to be able to perform analysis. Daytona simply will not work on data positioned elsewhere. Apache Spark is thus relatively flexible as it works with any set of features.

5.1. The initiatives that are based on MapReduce

Many BDA initiatives are based on MapReduce because of its popularity. The most common initiatives are explained below.

Haloop: Haloop is a modified version of the original MapReduce model. The modifications ensured that Haloop can perform iteration and related tasks [58]. Haloop works well with numerous ML tools [59] and can be used to develop CB-FPMs.

Twister: Twister works like Haloop. It is a light MapReduce runtime which improves MapReduce capability of supporting iteration tasks [60]. The improvement basically has to do with helping MapReduce perform faster on iteration tasks, making it viable for CB-FPM development. It works well with a number of ML tools and is operable both on cloud and on a cluster of computers [60,61].

Microsoft's Project Daytona: Like Twister, Daytona is also MapReduce runtime [62] that supports iterative computing. It is particularly designed to operate only on Microsoft Azure which is a free cloud platform that allows developing, organization and administration of applications. Daytona's inability to operate on other data sources/bases happens to be its major limitation. However, its special relations with Azure allows efficient performance by using Azure as the data source as well as data destination during computations [63]. Daytona requires no distributed file system to operate.

5.2. The initiative based on Spark

Apache Spark is an efficient and effective substitute to MapReduce. Spark uses a construct called Resilient Distributed Datasets (RDDs) that resourcefully supports machine learning problems that involve iterations [64]. As against MapReduce process which reads data from and writes results to a distributed file system for every iteration, Spark's RDDs help to keep data in memory for iterations until computation is completed thus increasing efficiency, speed and performance [64]. Speculations are that RDDs make Spark as many as 100 times faster than MapReduce in multi-pass analytics. Apache Mahout is the lone initiative on Spark.

Apache Mahout is a free scalable Machine learning tools library in BDA ecosystem. Mahout supports numerous ML tools (e.g. support vector machine, artificial neural network, among others) in executing clustering, filtering and classification analysis on huge data on a cluster of computers, cloud or a standalone computer [65]. Mahout used to be Hadoop based, using MapReduce model and consequently supported only ML tools that performed linear classifications e.g. linear support vector machine [66]. Presently a Spark base model, Mahout is now way more efficient and flexible as previously explained, and is fault tolerant. Table 1 is presentation of the requirements and/or features of the discussed initiatives while Fig. 4 presents a framework model that can be used to select an initiative for a problem.

6. Proposed framework architecture for construction firms failure prediction using big data analytics

Considering the highlighted potentials and challenges of CB-FPM development with BDA, a framework architecture for developing a BDA CB-FPM is proposed in Annex 1 figure. The framework

Table 1
Features of Big data initiatives capable of building FPM for construction firms.

Big data analytics initiative	Type/processing systems	Implementation platform	Distributed file system (Data Access)	Single or cluster/cloud	Support ML tools	Fault tolerance (FT)
Old Apache Mahout	MapReduce	Hadoop platform	HDFS	Both	Yes[a]	Yes
Daytona	MapReduce runtime	Microsoft Azure	Not required[b]	Cloud based only	Yes	Yes
Twister	MapReduce runtime	Twister platform	Twister tool[d]	Both	Yes	No[c]
Haloop	Modified Hadoop MapReduce	Hadoop Platform	HDFS	Cluster/cloud only	Yes	Yes
New Apache Mahout	Spark	Any platform[e]	Any system	Both	Yes	Yes

[a] Supports linear computation only.
[b] Microsoft Azure cloud provides a distributed file system by default.
[c] It is not fault tolerant for iterations.
[d] Twister provides a tool which manages data across distributed disks [60].
[e] Stand alone will require a distributed file system e.g. NFS mounted at the same path on each node.

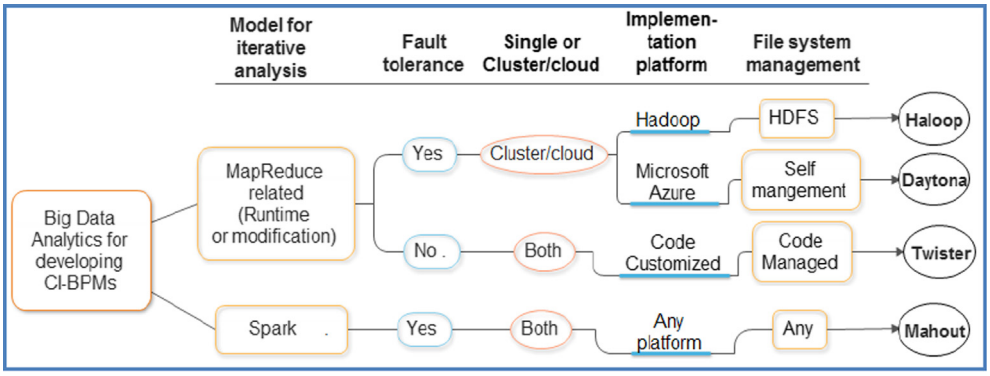

Fig. 4. A framework model for selection of suitable BDA initiative for developing CB-FPM. Self-management: The platform takes on the job of managing the file system. Code managed: The nodes need input code in order to manage file system. Any (file management system): This initiative is compatible with all file management systems.

Please cite this article in press as: H.A. Alaka et al., Applied Computing and Informatics (2018), https://doi.org/10.1016/j.aci.2018.04.003

begins with construction firms' data collection from databases into a single computer or a cluster of computers. This is followed by conversion of the data into the Key-Value Pair structure as sometimes required [54]. The required platform for the selected BDA initiative is installed. The data is then deposited on the conforming distributed file system. 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 BDA CB-FPM.

7. Framework implementation to develop BDA CB-FPM

7.1. The data

The data of tens of thousands of construction firms were painstakingly downloaded from FAME Bureau Van Dijk. As a test of the proposed BDA CB-FPM framework, data of 30,000 construction firms was extracted from the downloaded data. The proportion of failed to healthy construction firms in the extracted data was 50–50 to avoid the unequal data dispersion problem [67]. To aid quick model development, the five financial ratios used by Altman (1968) were used in this study, as done in some CB-FPM studies like Horta & Camanho [48], leading to 150,000 datacells. The ratios are as listed below

V1 = Working capital/Total assets
V2 = Retained Earnings/Total assets
V3 = Earnings before interest and taxes/Total assets
V4 = Market value equity/Book value of total debt
V5 = Sales (contacts values)/Total assets

ANN was the choice ML tool for this study because it usually requires a large number of iterations to achieve convergence, thereby causing long duration complex computations. The data was loaded onto two different computers. One was used to develop the CB-FPM without BDA while the other was set up to use BDA to develop the CB-FPM.

7.2. The big data analytics path selected and its set up

The path on the second row, going through Data Cloud to 'Reliable CB-FPM using Mahout', in the figure in Annex 1 was used to develop the BDA CB-FPM in this study. This path was chosen because of the flexibility of Apache Spark. The Amazon Elastic Compute Cloud (EC2) was used because of its cheap server compute 'Instances' which are capable of running applications. Although we had 15 Instances subscribed for, only ten were used for this experiment. The spark-ec2 was used to launch the 10 Instances. This ensured that the Apache Spark and HDFS on were automatically set up on the Instances. With the Instances running as nodes, one Instance was set as master node in the test-master group and the remaining nine as slave nodes in the test-slaves group.

The 'R' Language was the preferred analytics software because it has an Apache Spark package, called 'SparkR', which makes it easy to implement Spark. The SparkR package installed on R and the R program was connected to Spark using the sparkR.session command. The 150,000 datacells were loaded into R as a standard .csv data frame before being converted to SparkDataFrame file system supported by Spark.

7.3. The models and the results

The data was divided into 70% and 30% for training and testing respectively using the sample.split command. On the computer without BDA set up, the model training happened in a second but did not achieve convergence in the ANN default setting of 100 iterations maximum, leading to a suboptimal model. The ANN parameters were continually tuned to allow higher number of iterations in order to achieve convergence but this caused the training process to consume a lot of time. After setting the iteration limit to a maximum of one million iterations to allow as many as required to achieve convergence, the computer did not complete the training in nine hours and the process had to be aborted. This problem was envisaged hence the point of setting up the BDA platform in the first place.

With iterations limit set to one million from the start, training on the same data was run on the computer set up for BDA and it took about eight seconds for the model to converge at 460,000 iterations, clearly indicating that BDA is useful in developing CB-FPM with large data. The result of the model performance on test data are given in Table 2.

To check the effect of the large sample size on the results, data of 200 (100 existing and 100 failed) construction firms were randomly extracted from the complete dataset of 30,000 construction firms. Using the same five variables as before, to avoid bias during comparison, a new CB-FPM was developed with ANN without using the BDA platform. The model was successfully developed in about 2 s after 100 iterations. To test the CB-FPM, two separate data of 1000 firms (500 existing and 500 failed in each case) were randomly extracted from the remaining 29,800 construction firms. The two datasets were labelled 1000a and 1000b. The CB-FPM was tested with these 2 datasets to allow a check of reliability of the result.

The results of the BDA CB-FPM were not disappointing (see Table 2) with the model having an overall accuracy of 82.95% on the test data. The receiver operator characteristic (ROC) curve is presented in Fig. 5. The area under the curve (AUC) value which is used to measure the overall performance of an FPM came up as 0.8815536, showing a good overall performance for the BDA

Table 2
Results of the BDA CB-FPM and the ordinary CB-FPM on test data.

Model	BDA CB-FPM	Ordinary CB-FPM on 1000a dataset	Ordinary CB-FPM on 1000b dataset
Accuracy	82.95%	70.1	70%
AUC	0.8815536	0.7206454	0.7154232

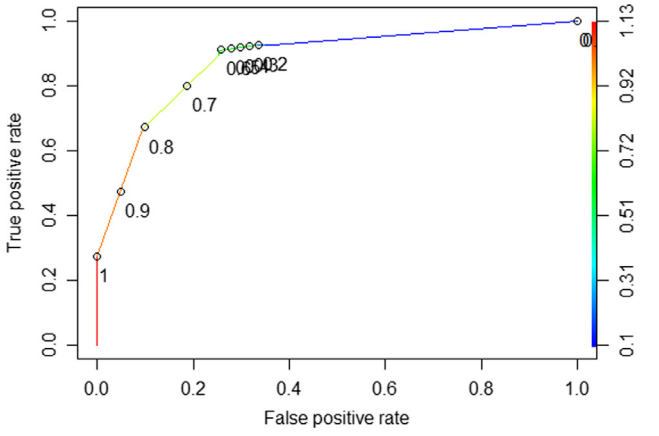

Fig. 5. Receiver operator characteristic (ROC) of the ANN CB-FPM developed with BDA.

CB-FPM (note that the closer the AUC value to one, the better, with 0.5 taken as the worst value). However, a careful variable selection process based on the sample construction firms' data used, an informed threshold modification away from the default 0.5, and further tuning of other ANN parameters like number of hidden nodes among others, could all have helped the model perform much better.

The results of the ordinary CB-FPM were poorer than those of the BDA CB-FPM (see Table 2) with the model having overall accuracies of 70.1% and 70% on the test data 1000a and 1000b respectively, compared to 82.95% of the BDA CB-FPM. To contextualise this, it is very important to understand that the cost of misclassifying a single construction firm can be devastating. A construction firm that is wrongly predicted as healthy when it is failing will cause the management team to carry on as normal, thereby causing the firm to eventually fail. Such failure will lead to financial loss for the firm's owner, job losses for the workers, revenue losses for office space renters, financial losses for owners of projects the firm is developing, various legal disputes, trauma for owner and workers, non-payment of suppliers among many other social and financial cost. The ROC curves of the ordinary CB-FPM performance on datasets 1000a and 1000b are presented in Figs. 6 and 7 respectively. The AUCs are much lesser than that achieved by the BDA CB-FPM (see Table 2), depicting a much lesser capability to perform well on new data.

The overall superiority of the BDA CB-FPM, and the associated reduction of great financial and social cost, shows that construction

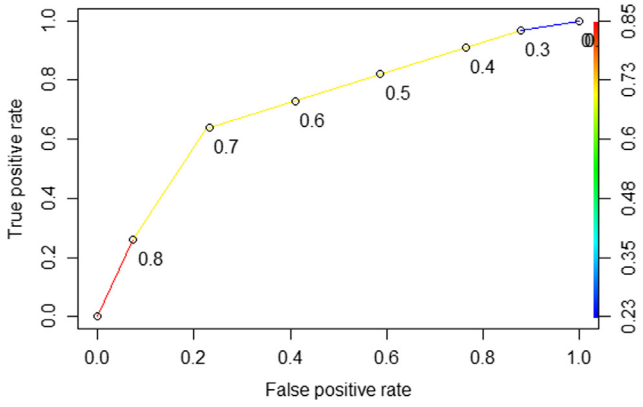

Fig. 6. Receiver operator characteristic (ROC) of the ordinary CB-FPM on test data 1000a.

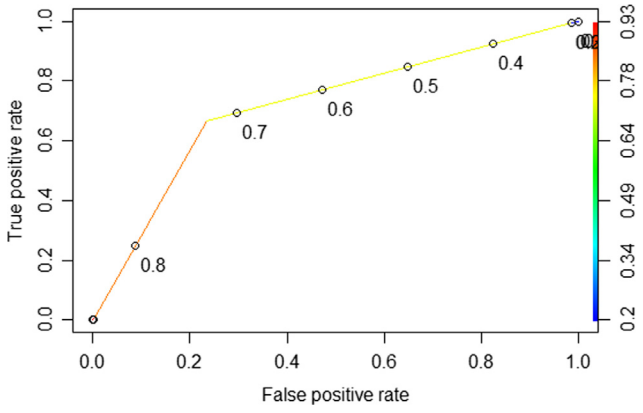

Fig. 7. Receiver operator characteristic (ROC) of the ordinary CB-FPM on test data 1000b.

Table 3
Influence the financial ratios based on the Information Gain method.

Variable codes	Variable names	Information gain values	Ranking
V1	Working capital/Total assets	0.31071137	1st
V5	Sales (contacts values)/Total assets	0.21711454	2nd
V4	Market value equity/Book value of total debt	0.20187683	3rd
V2	Retained Earnings/Total assets	0.20813130	4th
V3	Earnings before interest and taxes/Total assets	0.17103339	5th

firm owners, financiers, government bodies and policy makers need to adopt them.

Following the development and test of the ordinary and BDA CB-FPMs, a quick analysis was run to check the influence the financial ratios used as predictor variables. The analysis was done with the 'information gain' selector algorithm. The result is presented in Table 3. The higher the information gain value of a variable, the more the positive contribution of that variable to the predictive power of the CB-FPM. It is not surprising to see working capital/total assets as the chief contributor since it has to do with liquidity of the firm. Construction firms are known to always need high liquidity if they are to keep their projects running [29,48,68,69]. Poor liquidity is what led to the recent failure of Carilion construction firm in the United Kingdom. Although, V4 (see Table 3), which has something to do with the firm's debt, is tightly associated with liquidity, it did not come out as the second most important variable as expected. Instead, V5, which basically measures the rate at which a firm wins contracts, took that position. This is not too surprising as general knowledge cannot always be used to decide the level of positive contribution of the variables to the CB-FPM hence the use of selector algorithms like 'information gain'.

8. Conclusion

This study aimed to propose a framework architecture for developing a BDA CB-FPM and implementing it, using data of tens of thousands of construction firms. The readily available nature of financial data of hundreds of thousands of construction firms made them the ideal variable choice. It was discovered that MapReduce, which is the traditional big data analyser, is not fit to develop BDA CB-FPM because of its poor support for iteration. Many BDA initiatives consequently developed to support iteration were highlighted to include Haloop, Daytona, Twister and Spark among others. Based on the support features of each initiative, a framework clearly showing the path through which a reliable CB-FPM could be developed with each initiative was proposed. With Spark emerging as the most flexible, one of its paths was adopted to develop a BDA CB-FPM in this study using 150,000 datacells from financial statements of 30,000 construction firms. Using ANN with maximum number of iteration set to one million, a normal computer was unable to develop a CB-FPM in over nine hours. With BDA, the CB-FPM was developed in about eight seconds, achieving convergence at 460,000 iterations. The BDA CB-FPM outperformed an ordinary CB-FPM developed with a dataset of 200 construction firms randomly extracted from the 30,000 used for the BDA CB-FPM. It can thus be concluded that the problem of not being able to use a large dataset to develop CB-FPM is resolvable by applying BDA to CB-FPM development and that the framework proposed is valid. It can also be concluded that the use of a larger sample size, with the aid of BDA, can lead to

be performance of CB-FPMs. Future studies should look to use data of much more construction firms since larger data improves reliability. Effort should also be made to try other paths and initiatives in the proposed framework, and use of other ML tools.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.aci.2018.04.003.

References

[1] B. Giddings, B. Hopwood, G. O'Brien, Environment, economy and society: fitting them together into sustainable development, Sustain. Dev. 10 (4) (Nov. 2002) 187–196.
[2] C. Ucha, Poverty in Nigeria: some dimensions and contributing factors, Glob. Major. E-J. 1 (1) (2010) 46–56.
[3] D. for B. I. and Skills, "c Analysis of the Sector, BIS/13/958," London, 2013.
[4] C. Rhodes, "Construction Industry: Statistics and Policy.," London, 2015.
[5] European Commission, "FWC Sector Competitiveness Studies N° B1/ENTR/06/054–Sustainable Competitiveness of the Construction Sector-Final report," 2011.
[6] Department for Business Innovation and Skills, "Industrial Strategy: Government and Industry in Partnership," London, 2013.
[7] Global Construction Perspectives and Oxford Economics, "Global Construction 2030. A Global Forecast for the Construction Industry to 2030," London, 2015.
[8] The Insolvency service, "Insolvency Statistics – April to June 2016 (Q2 2016)," London, 2016.
[9] Dun and Bradstreet Limited, "Global Business Failures Report," 2012.
[10] M. Hodgson, "Insolvency figures show 23% of failures come from Construction Industry - Tremark," 2013. [Online]. Available: <http://www.tremark.co.uk/177-insolvency-figures-show-23-of-failures-come-from-construction-industry/> (Accessed: 05-Jan-2016).
[12] R.A.I. Van Frederikslust, Predictability of Corporate Failure, Springer, US, Boston, MA, 1978.
[13] E.I. Altman, G. Marco, F. Varetto, Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience), J. Bank. Financ. 18 (3) (1994) 505–529.
[14] J.-H. Chen, Developing SFNN models to predict financial distress of construction companies, Exp. Syst. Appl. 39 (1) (2012) 823–827.
[15] M. D. Odom and R. Sharda, "A neural network model for bankruptcy prediction," in: 1990 IJCNN International Joint Conference on Neural Networks, 1990, pp. 163–168 vol.2.
[16] W. Beaver, Financial ratios as predictors of failure, J. Account. Res. 4 (1966) (1966) 71–111.
[17] E. Altman, The success of business failure prediction models: an international survey, J. Bank. Financ. 8 (1984) 171–198.
[18] H. Fadel, The predictive power of financial ratios in the British construction industry, J. Bus. Financ. Account. 4 (3) (1977) 339–352.
[19] M.A. Carpenter, M.A. Geletkanycz, W.G. Sanders, Upper Echelons research revisited: antecedents, elements, and consequences of top management team composition, J. Manage. 30 (6) (2004) 749–778.
[20] B.R. Kangari, F. Farid, Financial performance analysis for construction industry, J. Constr. Eng. Manage. 118 (2) (1992) 349–361.
[21] D. Langford, R. Iyagba, D. Komba, Prediction of solvency in construction companies, Constr. Manage. 1993 (11) (1993) 317–325.
[22] A.F. Abidali, F. Harris, A methodology for predicting company failure in the construction industry, Constr. Manage. Econ. 13 (3) (1995) 189–196.
[23] J.S. Russell, H. Zhai, Predicting contractor failure using stochastic dynamics of economic and financial variables, J. Constr. Eng. Manage. 122 (2) (1996) 183–191.
[24] H.A. Alaka, L.O. Oyedele, H.A. Owolabi, V. Kumar, S.O. Ajayi, O.O. Akinade, M. Bilal, Systematic review of bankruptcy prediction models: towards a framework for tool selection, Exp. Syst. Appl. 94 (2018) 164–184.
[25] D. Singh, R.L.K. Tiong, Evaluating the financial health of construction contractors, Proc. Inst. Civ. Eng. – Munic. Eng. 159 (3) (2006) 161–166.
[26] Y. Huang, Prediction of contractor default probability using structural models of credit risk: an empirical investigation, Constr. Manag. Econ. 27 (6) (2009) 581–596.
[27] T. Sueyoshi, M. Goto, DEA–DA for bankruptcy-based performance assessment: Misclassification analysis of Japanese construction industry, Eur. J. Oper. Res. 199 (2) (2009) 576–594.
[28] H.P. Tserng, G.F. Lin, L.K. Tsai, P.C. Chen, An enforced support vector machine model for construction contractor default prediction, Autom. Constr. 20 (8) (2011) 1242–1249.
[29] S.T. Ng, J.M.W. Wong, J. Zhang, Applying Z-score model to distinguish insolvent construction companies in China, Habitat Int. 35 (4) (2011) 599–607.
[30] E. Makeeva, E. Neretina, The prediction of bankruptcy in a construction industry of Russian Federation, J. Mod. Account. Audit. 9 (2) (2013) 256–271.
[31] M. Muscettola, Probability of default estimation for construction firms – ProQuest, Int. Bus. Res. 7 (11) (2014) 153–164.
[32] H.P. Tserng, P.-C. Chen, W.-H. Huang, M.C. Lei, Q.H. Tran, Prediction of default probability for construction firms using the logit model, J. Civ. Eng. Manage. 20 (2) (2014) 247–255.
[33] J. Sun, B. Liao, H. Li, AdaBoost and bagging ensemble approaches with neural network as base learner for financial distress prediction of Chinese construction and real estate companies, Recent Patents Comput. Sci. 1932 (2013) 47–59.
[34] M.-Y. Cheng, N.-D. Hoang, L. Limanto, Y.-W. Wu, A novel hybrid intelligent approach for contractor default status prediction, Knowl. – Based Syst. 71 (2014) 314–321.
[35] J. Heo, J.Y. Yang, AdaBoost based bankruptcy forecasting of Korean construction companies, Appl. Soft Comput. 24 (2014) 494–499.
[36] K.S. Button, J.P.A. Ioannidis, C. Mokrysz, B.A. Nosek, J. Flint, E.S.J. Robinson, M.R. Munafò, Power failure: why small sample size undermines the reliability of neuroscience, Nat. Rev. Neurosci. 14 (5) (May 2013) 365–376.
[37] P. Du Jardin, Predicting bankruptcy using neural networks and other classification methods: the influence of variable selection techniques on model accuracy, Neurocomputing 73 (10) (2010) 2047–2060.
[38] P. Zikopoulos, C. Eaton, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, McGraw-Hill Osborne Media, New York, 2011.
[39] C.-W. Lee, K.-Y. Hsieh, S.-Y. Hsieh, H.-C. Hsiao, A dynamic data placement strategy for hadoop in heterogeneous environments, Big Data Res. 1 (2014) 14–22.
[40] Hadoop, "Welcome to ApacheTM Hadoop®!," 2014. [Online]. Available: <http://hadoop.apache.org/> (Accessed: 26-Feb-2015).
[41] J. Dean, S. Ghemawat, MapReduce, Commun. ACM 51 (1) (Jan. 2008) 107.
[42] Y. Li, K. Gai, L. Qiu, M. Qiu, H. Zhao, Intelligent cryptography approach for secure distributed big data storage in cloud computing, Inf. Sci. (NY) 387 (2017) 103–115.
[43] K. Gai, M. Qiu, H. Zhao, and J. Xiong, "Privacy-Aware Adaptive Data Encryption Strategy of Big Data in Cloud Computing," in: 2016 IEEE 3rd International Conference on Cyber Security and Cloud Computing (CSCloud), 2016, pp. 273–278.
[44] A. Jacobs, The pathologies of big data, Commun. ACM 52 (8) (Aug. 2009) 36.
[45] O. Bracht, "Five ways to handle Big Data in R | R-bloggers." WordPress, 2013.
[46] R.J. Mason, F.C. Harris, Predicting company failure in the construction industry, Proc. Inst. Civ. Eng. 66 (1979) 301–307.
[47] H.L. Chen, Model for predicting financial performance of development and construction corporations, J. Constr. Eng. Manage. 135 (11) (Nov. 2009) 1190–1200.
[48] I.M. Horta, a.S. Camanho, Company failure prediction in the construction industry, Exp. Syst. Appl. 40 (16) (Nov.2013) 6253–6257.
[49] E. Altman, Financial ratios, discriminant analysis and the prediction of corporate bankruptcy, J. Finance 23 (4) (1968) 589–609.
[50] R. Kangari, Business failure in construction industry, J. Constr. Eng. Manage. 114 (2) (1988) 172–190.
[51] S.S. Haykin, Neural Networks: A Comprehensive Foundation, Prentice Hall, New Jersey, 1994.
[52] S. Kale, D. Arditi, Age-dependent business failures in the US construction industry, Constr. Manage. Econ. 17 (4) (Jul. 1999) 493–503.
[53] D. Arditi, A. Koksal, S. Kale, Business failures in the construction industry, Eng. Constr. Archit. Manage. 7 (2) (2000) 120–132.
[54] A. Hafiz, O. Lukumon, B. Muhammad, A. Olugbenga, O. Hakeem, A. Saheed, "Bankruptcy Prediction of Construction Businesses: Towards a Big Data Analytics Approach," in: 2015 IEEE First International Conference on Big Data Computing Service and Applications, 2015, pp. 347–352.
[55] S. Madden, "From Databases to Big Data.," IEEE Internet Comput., vol. 16, no. 3, 2012.
[56] W. Fan, A. Bifet, Mining big data, ACM SIGKDD Explor. Newsl. 14 (2) (2013) 1.
[57] S. Wei and Z. Lin, "Accelerating Iterations Involving Eigenvalue or Singular Value Decomposition by Block Lanczos with Warm Start," Zhejiang, 2010.
[58] Y. Bu, B. Howe, M. Balazinska, M.D. Ernst, HaLoop, Proc. VLDB Endow. 3 (1–2) (2010) 285–296.
[59] V.S. Agneeswaran, Big Data Analytics: Evolution of Machine Learning Realizations, Pearson Education, New Jersey, 2014.
[60] J. Ekanayake, H. Li, B. Zhang, T. Gunarathne, S.-H. Bae, J. Qiu, and G. Fox, "Twister," in: Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing - HPDC '10, 2010, p. 810.
[61] S. Zhanquan, G. Fox, "Large Scale Classification Based on Combination of Parallel SVM and Interpolative MDS." 2012.
[62] H. Lei, T. Xing, J. Taylor, X. Zhou, Monitoring travel time reliability from the cloud, Transp. Res. Rec. J. Transp. Res. Board 2291 (2012) 35–43.
[63] R.S. Barga, J. Ekanayake, W. Lu, "Project Daytona: Data analytics as a cloud service," in: Proceedings – International Conference on Data Engineering, 2012, pp. 1317–1320.
[64] M. Zaharia, M. Chowdhury, M. Franklin, S. Shenker, "Spark: cluster computing with working sets.," in: HotCloud, 2010.
[65] Mahout, "Apache Mahout: Scalable machine learning and data mining," 2015. [Online]. Available: <https://mahout.apache.org/> (Accessed: 05-Nov-2015).

[66] K. Ericson, S. Pallickara, On the performance of high dimensional data clustering and classification algorithms, Futur. Gener. Comput. Syst. 29 (2013) 1024–1034.
[67] H.A. Alaka, L.O. Oyedele, H.A. Owolabi, S.O. Ajayi, M. Bilal, O.O. Akinade, Methodological approach of construction business failure prediction studies: a review, Constr. Manage. Econ. 34 (11) (2016) 808–842.
[68] H.A. Alaka, L.O. Oyedele, H.A. Owolabi, A.A. Oyedele, O.O. Akinade, M. Bilal, S.O. Ajayi, Critical factors for insolvency prediction: towards a theoretical model for the construction industry, Int. J. Constr. Manage. 17 (1) (2017) 25–49.
[69] H.A. Alaka, L.O. Oyedele, H.A. Owolabi, M. Bilal, S.O. Ajayi, O.O. Akinade, Insolvency of small civil engineering firms: critical strategic factors, J. Prof. Issues Eng. Educ. Pract. 143 (3) (2017) 4016026.

Please cite this article in press as: H.A. Alaka et al., Applied Computing and Informatics (2018), https://doi.org/10.1016/j.aci.2018.04.003