Bankruptcy prediction of construction businesses: Towards a big data analytics approach

Alaka, H, Oyedele, LO, Bilal, M, Akinade, OO, Owolabi, HA & Ajayi, SO

Original citation & hyperlink:
Alaka, H, Oyedele, LO, Bilal, M, Akinade, OO, Owolabi, HA & Ajayi, SO 2015, Bankruptcy prediction of construction businesses: Towards a big data analytics approach. in *2015 IEEE First International Conference on Big Data Computing Service and Applications (BigDataService).* IEEE, pp. 347-352, 2015 IEEE First International Conference on Big Data Computing Service and Applications (BigDataService), Redwood City, United States, 30-2 April
https://dx.doi.org/10.1109/BigDataService.2015.30

DOI 10.1109/BigDataService.2015.30
ESBN 978-1-4799-8128-1

Publisher: IEEE

BANKRUPTCY PREDICTION OF CONSTRUCTION BUSINESSES: TOWARDS A BIG DATA ANALYTICS APPROACH

Abstract

Bankruptcy prediction models (BPMs) are needed by financiers like banks in order to check the credit worthiness of companies. A very robust model needs a very large amount of data with periodic updates (i.e. appending new data). Such size of data cannot be processed directly by the tools used in building BPMs; however Big Data analytics offers the chance to analyse such data. With data sources like DataStream, FAME, Company House, etc. that hold large financial data of existing and failed firms, it is possible to extract huge financial data into Hadoop database (e.g. HBase), whilst allowing periodic appending of data from the data sources, and carry out a Big Data analysis using a machine learning tool on Apache Mahout. Lifelong machine learning can also be employed in order to avoid repeated intensive training of the model using all the data in the Hadoop database. A framework is thus proposed for developing a Big Data Analytics BPM

1.0 Introduction

The ferocity of business failures in the likes of the construction and manufacturing industries has led to continuous advancement in bankruptcy prediction research, usually in the form of developing bankruptcy prediction models (BPMs) using various tools such as artificial neural networks, rough set, etc. The use of BPMs to identify potential business failure can help prevent failures as well as ensure credit or contracts are given only to healthy companies. The performance of a BPM is dependent on the tool employed to build the model and the size and type of data available among other factors. To improve a BPM's overall performance and reliability, including the highly needed early prediction ability, very large data over a long period of time is needed with continuous update (i.e. appending new data to the old data) of such data (Haykin, 1994; Min and Lee, 2005; Tseng and Hu, 2010). Data, in this case, is usually in the form of periodic financial statement of firms providing financial ratios as variables. Although a few studies have attempted to use a really large set of data (Van Frederikslust, 1978; Altman, et al., 1994; Chen, 2012), the limitation on large data size handling and analysis has meant that no dataset has surpassed 10,000 in quantity and no prediction model or system has proposed automatic periodic update of the model by appending more dataset. This limitation can be eliminated with the use of Big Data Analytics which can in fact, be used to build a very robust model since it can analyse much more data and allows continuous appending of data.

Some studies have attempted to use a large dataset to develop BPMs for performance improvement purposes. Van Frederikslust's (1978), for instance, used the data of 40 (20 failed and 20 non-failed) sample firms over a period of 20 years leading to a relatively large data set of almost 800 annual financial statements. Altman, et al. (1994) did much better by using data of 1000 firms over a 10 year period to generate almost 10,000 financial reports dataset to build their model. Recently, Chen (2012) also generated 1615 financial reports from 42 sample construction firms. Although these are relatively large datasets compared to other BPM studies, they are too small to build a robust BPM. The authors might have well been conscious and cautious of the fact that none of the available tools can analyse the huge size of data required for robust BPM. This also meant that continuous appending (update) of new data to the model was ruled out since it will increase the data size.

Since an increase in data size leads to increased performance and reliability of BPMs, to develop a robust prediction system, for example, data for all construction companies in the US (currently 729,345) over the past 20 years can be used to develop a prediction system which will allow continuous periodic appending of data of new and existing firms. With so many failed firms in the past 20 years, over one million companies will be involved in the sample firms and over 15 million financial statements will be ready as the initial data, with more data to be appended to the system periodically. The management and analysis of such large volume and dynamic data will require a Big Data Analytics approach. As such, the aim of this study is to explore how Big Data can be used to develop an updatable and append-able robust failure prediction system for the construction industry of any country with large records of data. The following objectives are required to achieve this aim:

- To identify potential data sources that contain financial statements of most construction firms in a country.
- To identify the tools that are commonly used to develop failure prediction models
- To explore how 'Big Data Analytics' can be used to store and process data from these data sources using common BPM tools in order to develop a robust prediction system

This study contributes to knowledge by being the first to explore how a robust BPM could be developed with huge financial data using Big Data technology. It proposes a framework for developing a Big Data Analytics BPM

The next section explains what Big Data is and the features that allows a data or dataset to be qualified as Big Data. Section 3 briefly highlights the popular and promising tools used in developing BPMs. Section 4 talks about the suitable variable type and potential sources of data for a BPM to be developed with Big Data technology. Section 5 explains what can be used to execute the Big Data Analytics of the financial data using any of the tools identified in order to develop a robust BPM and proposes a model of the whole process. Section 6 is a conclusion to the study.

2.0 Big Data Analytics and the Suitability of Bankruptcy Prediction Data

According to Diebold (2012), the combination 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". Although it is hard to give a precise definition of Big Data since 'big' as a word is a relative one, the concept of Big Data can be said to deal with three main characteristics of data namely: volume, variety and velocity (Zikopoulos and Eaton, 2011). Volume deals with size of the data, velocity deals with the rate at which data is being generated and need to be utilized while variety is concerned with the degree of variability of the data (Zikopoulos and Eaton, 2011). Big data mostly has to do with unstructured data (Suthaharan, 2014). Contrary to popular belief however, structured data can also be classified as Big Data and analysed with Hadoop depending on other features of the data (Zikopoulos and Eaton, 2011).

A data is considered as Big Data when its volume, variety and velocity become so high that present technology finds it hard to store and/or process (Pflugfelder, 2013; Suthaharan, 2014). It is a data size that compels a request for approaches other than the tried and trusted methods. In the 1980s, it could have been data which required 'tape monkeys'; at present, it is data that requires massively parallel programs running on a number of servers (Fan and Bifet, 2013). Big data analytics is defined as involving analysis of huge data in order to unmask valuable patterns/information (Suthaharan, 2014).

Although the size of a data is key to qualify it as Big Data, the type of analysis is as important. Jacobs (2009) experimented with why a data with basic demographic information (age, sex, ethnicity etc,) of the world population which would create a table of over 7.1 billion rows, about ten columns and fit into a 100 gigabyte disk should be classified as Big Data or not. 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 this data on this database will obviously take days if not weeks or months hence it can be classified as Big Data in this case.

The above example is what makes the data of hundreds of thousands of construction companies in a large country (e.g. USA, China, India, etc.) or region (Europe) over a number of years qualify as Big Data. A simple input of such data into columns and rows of Microsoft word and finding averages might not be considered as ‘Big’; however, a more complex analysis (like classification analysis which is used for bankruptcy prediction) of such a huge data using an machine learning (ML) tool will be nearly impossible on any computer. Such analysis hence qualify the data for Big Data Analytics.

Data warehouse is sometimes seen as the solution to the analysis of huge data in business applications. A warehouse of data can be defined as “a copy of transaction data specifically structured for query and analysis,” (Kimbal, 1996). It starts with mass data extraction before reconstituting in a separated database in a way that is easier to analyse (Jacobs, 2009). However huge data with continuous periodic appending, as is the case with financial statements of many firms which are submitted periodically, cannot be handled by data warehouse when it involves complex analysis as it is with developing BPMs (Jacobs, 2009; Madden, 2012). This reinforces the need for Big Data Analytics in this case.

3.0 Tools used for Developing Bankruptcy Prediction Models

Bankruptcy prediction is a classification problem which requires firms to be classified as failing or non-failing and tools used to create BPMs can be statistical or ML tools. The most popular statistical tools include multi-discriminant analysis (MDA) and logit analysis (LA) while popular and promising AI tools include artificial neural networks (ANN), support vector machines (SVM), rough sets (RS), case based reasoning (CBR), iterative dichotomiser 3 (ID3) and genetic algorithm (GA). A brief description of each tool is given in table 1.

Table 1: Working principles and details of tools used to develop BPMs

Tool	Category	Principle	References
Multi-discriminant analysis	Statistical	MDA produces a discriminant function with assigned coefficients to selected variables. The coefficients do not represent the level of importance of each variable. The function is used to calculate a single value known as Z-score. A cut-off score is chosen to classify firms as failed or non-failed based on the s sample firms	Altman (1968)
Logit analysis	Statistical	Like MDA, LA produces a logistic function with assigned coefficients to selected variables. The coefficients in this case however represent the level of importance of each variable. The function is used to calculate a binary score i.e. 0 or 1. One of the scores represents failing firm and the other represents non-failing firm.	Ohlson (1980); Jackson and Wood (2013)
Artificial neural networks	Machine learning	This tool imitates the brain's neural system in order to make classifications. A set of sample is used to train the network before the network is able to perform classifications. Overtraining or undertraining can lead to low performing models. Classification output is given with a binary score	Hertz et al. (1991); Jo and Han (1996); Chung et al. (2008)
Support vector machines	Machine learning	SVM constructs a dividing hyperplane on the selected variables of sample firms which separates the firms into failing and non-failing. The variables SVM finally employs for classification are only those that are close to the separating hyperplane. Classification output is given with binary score.	Hearst et al. (1998); Dreiseitl and Ohno-Machado (2002); Shin et al. (2005)
Rough sets	Machine learning	This tool operates by assuming all objects have attributes that define them. It constructs a partition between objects (i.e. sample firms) by grouping objects with common or similar attributes together. It then extracts decision rules for classification	Pawlak (1982); Ravi Kumar and Ravi (2007); Greco et al. (2001)

Case based reasoning	Machine learning	CBR stores cases (i.e. sample firms in this case) in a case library from which classification is made based on how closely a firm's attributes are to those of a sample firm in the case library. CBR is very easy to update; new cases are simply added to its library. It extracts decision rules for classification.	Kolodner (1993); Jo and Han (1996); Shin and Lee (2002)
Iterative dichotomiser 3	Machine learning	ID3 first determines the most discriminating variables between failing and non-failing firms in a sample set and then constructs a recursive partition between the firms. It subsequently extracts decision rules , based on the partition, for classification	Quinlan (1986); Tam and Kiang (1992); Anyanwu and Shiva (2009).
Genetic algorithm	Machine learning	GA is an optimizing search tool which identifies the global minimum in a search space by imitating the Darwin's evolution principle. It extracts decision rules for classification. All the rules extracted must be satisfied before a decision can be made	Shin and Lee (2002); Ravi Kumar and Ravi (2007)

4.0 Variables, Potential Data Sources and Challenges

The variables used to develop bankruptcy prediction models can be quantitative or qualitative. Quantitative usually come in form of financial ratios while qualitative are usually gotten from literature or enquiries (questionnaires, case study, etc.). For readily available data as needed in the case of Big Data Analytics which requires the data to keep coming, only financial ratios as quantitative variables will be viable. They are overall more commonly employed for developing BPMs because of their readily available nature (Balcaen and Ooghe 2006). The number of financial ratios to be used as variables does not need to be limited as is commonly done for normal BPMs since Big Data Analytics can handle large data. This can ensure a BPM is even more robust since most BPM studies have selected different financial ratios as being the most important for developing a BPM; proving that most ratios are important.and being able to use them all can improve reliability. The number of ratios that can be used will however depend on the number available in the data sources.

The sources for financial information (or financial ratios) for public companies are quite simple to identify. A data source like DataStream[1] hosts the financial information of many public companies around the world (including US and Europe). On the other hand, FAME[2] **(**Financial Analysis Made Easy) has financial information on over 9 million existing public and private companies and 5 million failed companies in UK. Financial information of over 3 million existing and a large number of failed non-listed companies in the UK can also be gotten from Company House[3] while such information on US non-listed companies can be gotten from state equivalent of Company House. Most of these data sources can export data direct to excel hence exporting the required data should not be a very big problem. However, converting the data to the Key-Value Pair format of a Big Data Analytics database/server (e.g. Hadoop HBase), where the Big Data Analytics can be carried out, might pose some challenges. The continuous appending of financial data to update the Hadoop database for a BPM will face the same challenges. There can also be a number of challenges to getting the financial data which can include legal and cost requirement for direct near-real time access.

Another set of challenges is 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 firms might be missing 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 BPMs. These challenges are however not a big problem since Big Data Analytics has the competence to analyse unstructured, uncertain and incomplete data. It might however be necessary to only append the model with data of up to about a year priori by always leaving out the data for the most recent year. This will ensure that the model that is built/appended does not contain the latest data of firms that are to be checked for potential bankruptcy since it is the latest financial statement that is used for this exercise; otherwise the model might not be effective for assessing any of the companies whose data is used to build it.

5.0 Bankruptcy Prediction Using Big Data and Machine Learning Tools

None of the tools used for building BPMs 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) hence using ML tools to directly analyse Big Data is virtually impossible (Fan and Bifet, 2013). One major problem of using ML tools for classification analysis on Big Data is that when a ML tool is trained with a specific data, it might not be suitable for another dataset (Suthaharan, 2014). Another problem is that while the ML tool will be trained to recognize mainly two classes (i.e. failed and non-failed firms), any other possible classes identified by the tool due to continuous data appending can lead to inaccuracies in classification. Further, very accurate tools like ANN and SVM have many parameters that increase their computational complexity making them unfit for Big Data Analytics. An attempt to reduce this computational complexity has led to development of different variations of tools like SVM (Giacinto et al., 2005; Laskov et al., 2004). Many research programmes have thus been designed to find a way data processing platforms like database management systems (DBMSs) (e.g. data warehousing) and MapReduce (Big Data Analytics), and packages/machine learning tools like R, Matlab, ANN, SVM etc. can work together (Madden, 2012).

5.1 Execution of machine learning on Big Data

The best option of platform to execute a Big Data BPM (classification) problem, using MapReduce, presently is the Apache Mahout which readily provides a structure on which numerous ML algorithms can be executed on top of MapReduce (Madden, 2012; Fan and Bifet, 2013). "*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). Hundreds of programs have been successfully executed using this model (Dean and Ghemawat, 2008). Other analysis type options include Microsoft's Project Daytona, University of California's (Berkeley) Spark, University of Washington's HaLoop and Indiana University's Twister (Madden, 2012). These platforms offer great support for certain ML tools in the MapReduce form but "still lack database systems' data management features" (Madden, 2012, p.6). Asides Mahout, another open source initiative platform is Massive Online Analysis (Bifet et al., 2010)

- Lifelong Machine Learning (LML) on Big Data

  One of the major challenges for the use of Big Data Analytics and ML tools for bankruptcy prediction is the continuous inflow of data in the form of financial ratios. When a new set of data is appended to a ML tool BPM, a retrain of the model is generally required to keep it updated. Continuous retrain of the whole huge data in a Big Data Analytics database however might not be a very good idea since the exercise can be really intensive. The probable solution to this challenge is LML (Silver and Poirier, 2007; Silver, 2011). This is because LML, after an initial training, has the ability to retain knowledge from the previous training and subsequently combine such knowledge with the knowledge acquired from training with a new data (Silver et al., 2013). In addition, it has the competence to use the retained knowledge to improve learning (training) on new data (Silver and Poirier, 2007). LML can thus help to eliminate the continuous intensive retraining of the Big Data BPM with full data. The implementation of LML on Big Data is however not seamless. One of the implementation challenges include the scalability which is an essential condition for Big Data applications (Silver, 2011). Another challenge is the continuous validation of the model whenever there is new data so that model is not unnecessarily retrained if it still suitable for the classification process (Suthaharan, 2014). This is not a big problem with bankruptcy prediction since the velocity of the incoming data is not extremely high and retraining might only be needed quarterly as against daily or weekly in some other applications.

- Proposed Framework for developing a Big Data Analytics BPM

Figure one presents a proposed framework for developing a Big Data Analytics BPM. In the framework, huge financial data of construction companies is extracted from numerous data sources and converted to the Key-Value Pair structure before being imported into a Big Data Analytics database such as HBase. Apache Mahout with LML is subsequently used to perform a classification analysis on the huge data using a BPM machine learning tool. This produces a classification result which predicts firms as either failing or non-failing. The LML is then employed for training every time new data is appended in order to avoid the intensive retraining using the full data

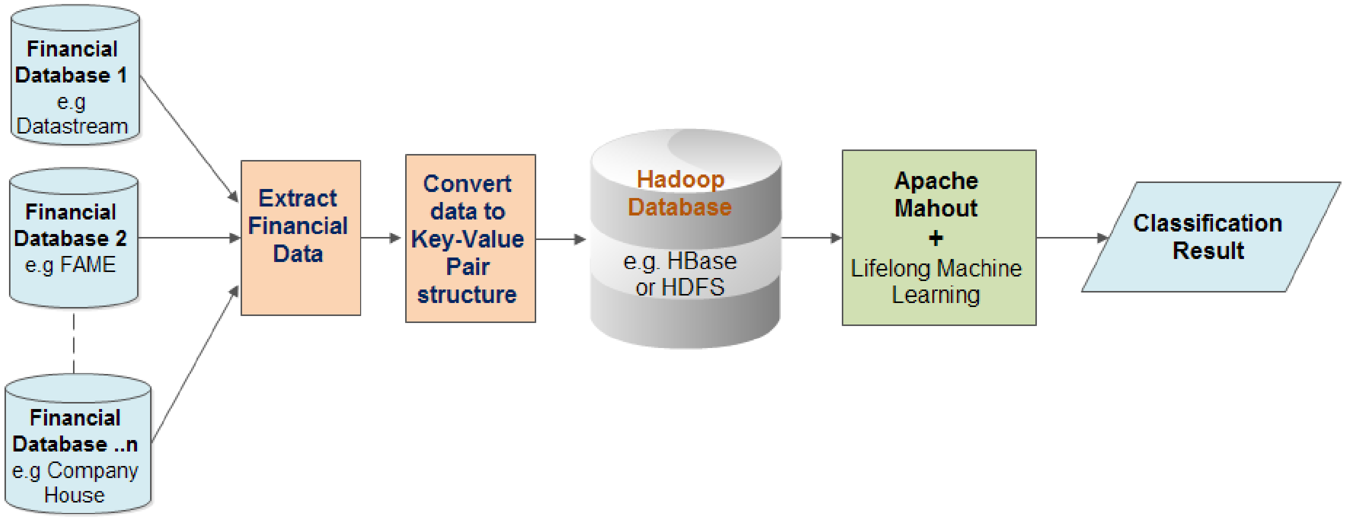

Figure 1: Flow diagram for application of Big Data Analytics on bankruptcy prediction.

6.0 Conclusion

The construction industry is one of the most vital to any country's economy despite its high record of business failure. Bankruptcy prediction for businesses in such industry, and any industry at all, is of paramount importance to owners in order to reduce business failures, and to financiers and clients in order to assess businesses for loans and contracts respectively. BPMs can hence help improve the economy. Despite reasonably extensive efforts, no BPM has been found to be robust enough usually because of the tool employed, insufficient data and/or updating/appending problems. Finally, Big Data Analytics offers the chance to build a robust model as it is capable of analysing all the available data in any data sources i.e. almost any size of data.

The data suitable for this type of model would be quantitative variables data in the form of financial ratios as against qualitative data or combined data. This is because they are readily available in data sources from where they can be extracted into a Big Data Analytics database (e.g. Hadoop's HBase). Unstructured, uncertainty and incomplete information of data from data sources is not a problem since Big Data Analytics is good for analysing unstructured data

The common ML tools used bankruptcy prediction models studies include ANN, SVM, RS, etc. Although these tools are unfit to directly analyse huge data, the Apache Mahout provides a suitable platform for some of these tools to use Big Data Analytics to solve classification problems. The problem of repeated intensive training on full data every time there is new data can be solved by using

LML. Overall, this work shows that Big data Analytics can be used for bankruptcy prediction by developing very robust BPMs. A framework is thus proposed for developing a Big Data Analytics BPM.

loud computing is a type of parallel distributed computing system that has become a frequently used computer application. MapReduce is an effective programming model used in cloud computing and large-scale data-parallel applications. Hadoop is an open-source implementation of the MapReduce model, and is usually used for data-intensive applications such as data mining and web indexing. The current Hadoop implementation assumes that every node in a cluster has the same computing capacity and that the tasks are data-local, which may increase extra overhead and reduce MapReduce performance. This paper proposes a data placement algorithm to resolve the unbalanced node workload problem. The proposed method can dynamically adapt and balance data stored in each node based on the computing capacity of each node in a heterogeneous Hadoop cluster. The proposed method can reduce data transfer time to achieve improved Hadoop performance. The experimental results show that the dynamic data placement policy can decrease the time of execution and improve Hadoop performance in a heterogeneous cluster.

Lee, C. W., Hsieh, K. Y., Hsieh, S. Y., & Hsiao, H. C. (2014). A Dynamic Data Placement Strategy for Hadoop in Heterogeneous Environments. *Big Data Research*, *1*, 14-22.

REFERENCES

Altman E.I. (1968). Financial ratios discriminant analysis and the prediction of corporate bankruptcy, *The Journal of Finance*, **23** (4), pp. 589-609.

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.

Anyanwu, M.N. and Shiva, S.G. (2009). Comparative analysis of serial decision tree classification algorithms. *International Journal of Computer Science and Security*, **3** (3), pp. 230-240.

Balcaen and Ooghe (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

Bifet, A., Holmes, G., Kirkby, R., and Pfahringer, B. (2010). Moa: Massive online analysis. *The Journal of Machine Learning Research*, *11*, 1601-1604.

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

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.

Dean, J., and Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM*, **51** (1), 107-113.

Diebold, F. (2012). On the Origin (s) and Development of the Term” Big Data. ”.*Penn Institute for Economic Research, Pier Working Paper*, pp. 12-037.

Dreiseitl, S., and Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics*, **35** (5), pp. 352-359

Fan, W., and Bifet, A. (2013). Mining big data: current status, and forecast to the future. *ACM SIGKDD Explorations Newsletter*, **14** (2), pp. 1-5.

Giacinto, G., Perdisci, R., and Roli, F. (2005). Network intrusion detection by combining one-class classifiers. In: *Image Analysis and Processing–ICIAP 2005* (pp. 58-65). Springer Berlin Heidelberg.

Greco, S., Matarazzo, B., and Slowinski, R. (2001). Rough sets theory for multicriteria decision analysis. *European Journal of Operational Research*, **129** (1), pp. 1-47.

Haykin, S. (1994). *Neural Networks: A Comprehensive Foundation*. New York: Mc Millan.

Hearst, M. A., Dumais, S. T., Osman, E., Platt, J., and Scholkopf, B. (1998). Support vector machines. *Intelligent Systems and their Applications, IEEE*, **13** (4), pp. 18-28.

Hertz, J., Krogh, A. and Palmer, R. (1991). *Introduction to the Theory of Neural Computing*. New York: Addison Wesley.

Jackson, R. H., 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. (2009). The pathologies of big data. *Communications of the ACM,* **52** (8), pp. 36-44.

Jo, H., and Han, I. (1996). Integration of case-based forecasting, neural network, and discriminant analysis for bankruptcy prediction. *Expert Systems with applications*, **11** (4), pp. 415-422.

Kimbal, R. (1996). *The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses*. New York: John Willey and Sons.

Kolodner, J. (1993). *Case-Based Reasoning*. San Francisco, California: Morgan Kaufmann Publishers Inc.

Laskov, P., Schäfer, C., Kotenko, I., and Müller, K. R. (2004). Intrusion detection in unlabeled data with quarter-sphere support vector machines. *Praxis der Informationsverarbeitung und Kommunikation*, **27** (4), 228-236.

Madden, S. (2012). From databases to big data. *IEEE Internet Computing*, **16** (3), pp. 4-6.

Min, J. H., and Lee, Y. C. (2005). Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert Systems with Applications*, **28** (4), pp. 603-614.

Ohlson, J.A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, **18** (1) pp. 109-131.

Pawlak, Z. (1982). Rough sets. *International Journal of Computer and Information Sciences*, **11** (5), pp. 341-356.

Pflugfelder, E. H. (2013). Big data, big questions. *Communication Design Quarterly Review*, **1** (4), pp. 18-21.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, **1** (1), pp. 81-106.

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.

Shin, K. S., and Lee, Y. J. (2002). A genetic algorithm application in bankruptcy prediction modeling. *Expert Systems with Applications*, **23** (3), pp. 321-328.

Shin, K. S., Lee, T. S., and Kim, H. J. (2005). An application of support vector machines in bankruptcy prediction model. *Expert Systems with Applications*, **28** (1), pp. 127-135.

Silver, D. L. (2011). Machine Lifelong Learning: Challenges and Benefits for Artificial General Intelligence. In: *4th International Conference, AGI 2011, Mountain View, CA, USA, August 3-6, 2011. Proceedings*, **6830** (2011), pp. 370-375. Heidelberg: Springer Berlin Heidelberg.

Silver, D. L., and Poirier, R. (2007). Requirements for machine lifelong learning. In: *Bio-inspired Modeling of Cognitive Tasks* (pp. 313-319). Heidelberg: Springer Berlin Heidelberg.

Silver, D. L., Yang, Q., and Li, L. (2013). Lifelong Machine Learning Systems: Beyond Learning Algorithms. In *AAAI Spring Symposium: Lifelong Machine Learning*.

Suthaharan, S. (2014). Big data classification: Problems and challenges in network intrusion prediction with machine learning. *ACM SIGMETRICS Performance Evaluation Review*, **41** (4), pp. 70-73.

Tam, K. Y., and Kiang, M. Y. (1992). Managerial applications of neural networks: the case of bank failure predictions. *Management science*, **38** (7), pp. 926-947.

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.

Van Frederikslust, R. A. I. (1978). *Predictability of Corporate Failure: Models for Prediction of Corporate Failure and for Evaluation of Debt Capacity*. Martinus Nijhoff Social Sciences Division.

Zikopoulos, P., and Eaton, C. (2011). *Understanding big data: Analytics for enterprise class hadoop and streaming data*. New York: McGraw-Hill Osborne Media.

1) http://financial.thomsonreuters.com/en/products/tools-applications/trading-investment-tools/datastream-macroeconomic-analysis.html

2) http://www.bvdinfo.com/en-gb/our-products/company-information/national-products/fame

3) https://www.gov.uk/government/organisations/companies-house/about/about-our-services