International Journal On Engineering Technology and Sciences – IJETS™
ISSN(P): 2349-3968, ISSN (O): 2349-3976
Volume I, Issue III, July - 2014

A DETAILED REVIEW ON DATA MINING IN FINANCE SECTOR

Dr.C.Kumar Charliepaul [1]
Principal
A.S.L Pauls College of Engg & Tech, Coimbatore .
charliepaul1970@gmail.com

G.Immanual Gnanadurai [2]
Assistant professor / CSE
Dhaya College of Engineering, Madurai
imman8601@gmail.com

Abstract: *Financial data is used in many financial institutes for precise analysis of consumer data to find debtor and legitimate customer. This data be capable of be stored and maintained to produce information and facts. This information and knowledge has to be circulated to every stake holders for the effective decision making process. Due to the enhance in the data, it is significant to mine knowledge/information from the large data repositories. Hence, Data mining has become an crucial factor in various fields including business, education, health care, finance, scientific etc., . This paper discusses about data mining and its techniques used in financial sector.*

Keywords: *Data mining, financial risks, Data mining techniques.*

1. Introduction

The intention of data mining is to mine valuable information from large databases or data warehouses. Data mining applications are used for business-related and scientific sides. This revise generally discusses the Data Mining applications in the scientific side. Scientific data mining distinguishes itself in the sense that the nature of the datasets is often very different from conservative market obsessed data mining applications. In this work, a detailed review is carried out on data mining in the finance sector, types of data used and details of the information extracted. Data mining algorithms applied in finance industry play a significant role. To find the valuable and concealed knowledge from the database is the function behind the application of data mining. Commonly data mining called knowledge discovery from the data.

1.1Need of Data Mining in Finance

Essentials of data mining in finance are upcoming from the need to:

- Predict *multidimensional time series* with high level of *noise;*
- Contain specific competence criteria (e.g., the maximum of trading Profit) in addition to prediction accuracy such as *R*2; make coordinated *multi-tire solution forecast* (minutes, days, weeks, months and years)
- incorporate a *stream of text signals* as input data for forecasting models
- able to *explain the forecast* and the *forecasting model* ("black box"
- models have limited interest and future for significant investment decisions);
- able to benefit from very *subtle patterns* with a *short life time;* and
- integrate the blow of market players on market regularities

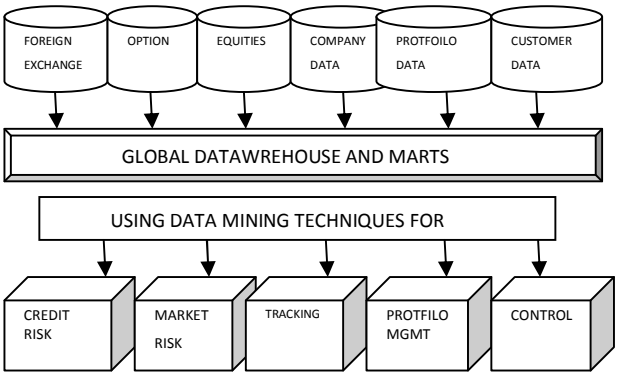


Fig 1: *The use of Data Mining Technique is a Global and Firm wide challenge for financial business. Firm-wide data source can be used through data mining for different business areas.*

2. Data Mining

Data Mining is the process of extracting knowledge hidden from large volumes of raw data. The knowledge must be new, not evident, and one must be able to use it. Data mining has been defined as

The practice of examining large pre-existing databases in order to generate new information [1]. It is "the science of extract useful information from large databases" [2, 3]. Data mining is one of the tasks in the procedure of knowledge discovery from the database [4]. Fig. 2 shows the process of knowledge discovery.

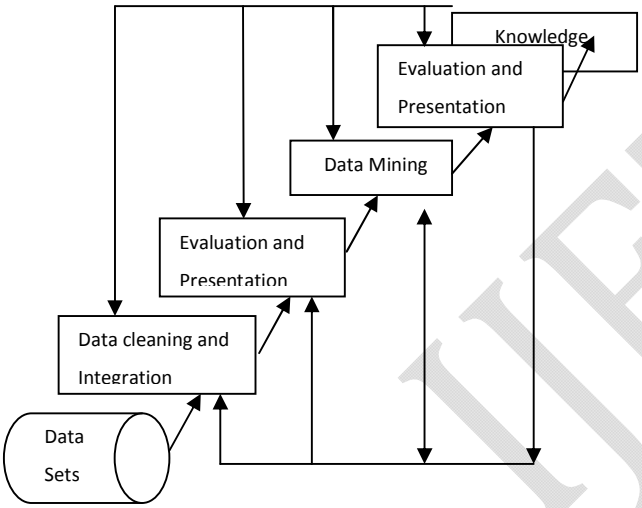


Fig 2.*Steps in KDD*

Data mining process can be broken down to the following iterative sequence of following steps.

1. Learning the application domain: includes relevant prior knowledge and the goals of the application.

2. Creating a target dataset: includes selecting a dataset on which discovery is to be performed.

3. Data cleaning and pre-processing: includes basic operations, such as removing noise or outliers if appropriate, collecting the necessary information to model or account for noise, deciding on strategies for handling missing data fields.

4. Data reduction and projection: includes finding useful features to represent the data, depending on the goal of the task, and using dimensionality reduction or transformation methods to reduce the effective number of variables under consideration or to find invariant representations for the data.

5. Choosing the function of data mining: includes deciding the purpose of the model derived by the data mining algorithm.

6. Choosing the data mining algorithm(s): includes selecting method(s) to be used for searching for patterns in the data, such as deciding which models and parameters may be appropriate and matching a particular data mining method with the overall criteria of the KDD process .

7. Data mining: includes searching for patterns of interest in a particular representational form or a set of such representations, including classification rules or trees, regression, clustering, sequence modeling, dependency, and line analysis.

8. Interpretation: includes interpreting the discovered patterns and possibly returning to any of the previous steps, as well as possible visualization of the extracted patterns, removing redundant or irrelevant patterns, and translating the useful ones into terms understandable by users.

9. Using discovered knowledge: includes incorporating this knowledge into the performance system, taking actions based on the knowledge.

3. Types of Risks Factors in Financial sector

3.1 Risk Management

Managing and measurement of risk is at the core of each financial institution. Today's major confront in the banking and insurance world is consequently the accomplishment of risk management systems in regulate to identify, measure, and control business exposure. Here credit and market risk current the central challenge, one can examine a major modify in the area of how to measure and deal with them, based on the advent of advanced database and data mining technology today, integrated measurement

of different kinds of risk (i.e., market and credit risk) is moving into focus. These all are based on models representing single financial instruments or risk factors, their behavior, and their interaction with overall market, making this field highly important topic of research [5].

3.2 Financial Market Risk

For single financial instruments, that is, stock indices, interest rates, or currencies, market risk extent is based on models depending on a set of underlying risk factor, such as interest rates, stock indices, or economic development. One is interested in a functional variety between instrument worth or risk and underlying risk factors as well as in functional dependence of the risk factors itself.

Today different market risk measurement approaches exist. All of them rely on models representing single instrument, their behavior and interaction with overall market. Many of this can only be built by using various data mining techniques on the proprietary portfolio data, since data is not publicly available and needs dependable supervision.

3.3 Credit Risk

Credit risk evaluation is key component in the procedure of commercial lending. Without it the lender would be unable to make an objective judgment of endure to lend to the prospective borrower, or if how much charge for the loan. Credit risk management can be classified into two basic groups:

a. Credit scoring/credit rating. Assignment of a customer or a product to risk level. (i.e., credit approval)
b. Behavior scoring/credit rating migration analysis. Valuation of a customer‘s or product’s probability of a change in risk level within a given time.(i.e., default rate volatility)

In commercial lending, risk assessment is usually an endeavor to quantify the risk of loss to the lender when making an exacting lending decision. Here credit risk can enumerate by the changes of value of a credit product or of a whole credit customer portfolio, which is based on change in the instrument’s ranting, the default probability, and recovery rate of the instrument in case of default. Further diversification effects authority the result on a portfolio level. Thus a major part of implementation and care of credit risk management system will be a typical data mining problem: the modeling of the credit instrument’s value throughout the default probabilities, rating migrations, and revival rates.

Three foremost approaches exist to model credit risk on the transaction level: accounting analytic approaches, statistical prediction and option theoretic approaches. Since large amount of information about client exist in financial business, an adequate way to assemble such models is to use their own database and data mining techniques, fitting models to the business needs and the business current credit portfolio.

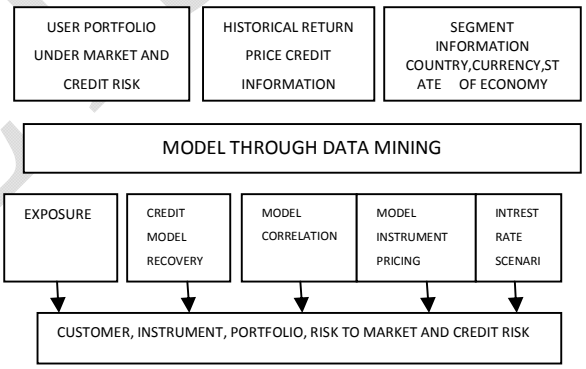


Fig 3: *Using Data mining technique for customer, financial instrument, portfolio risk to market and credit risk measurement*

3.4 Portfolio Management

Risk measurement approaches on an aggregated portfolio level quantify the risk of a set of instrument or customer counting diversification effects. On the other hand, forecasting models give a commencement of the expected return or price of a financial instrument. Both make it possible to manage firm wide portfolio actively in a risk/return efficient manner. With the data mining and optimization technique investors are able to allocate capital across trading activities to maximize profit or minimize risk. This

characteristic supports the ability to generate trade suggestion and portfolio structuring from user supplied profit and risk prerequisite. With data mining techniques it is possible to provide general situation scrutiny capability regarding expected asset prices or returns and the risk involved. With this functionality, what if simulations of varying market situation e.g. interest rate and exchange rate changes) cab be run to assess impact on the value and/or risk associated with portfolio, business unit counterparty, or trading desk. Various scenario results can be regard by consider actual market conditions. Profit and loss analyses allow user to entrance an asset class, region, counterparty, or custom sub portfolio can be benchmarked against familiar international benchmarks.

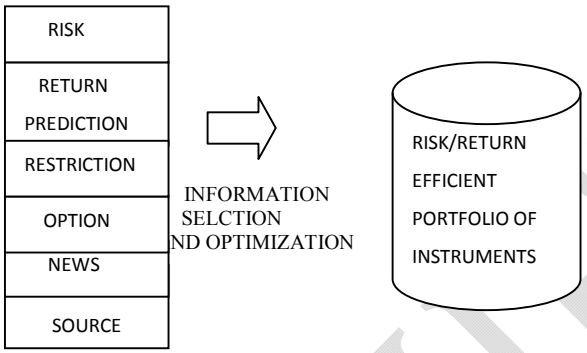
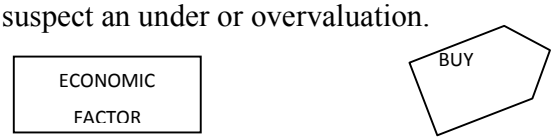


Fig 4: *The management of an instrumental portfolio is based on all reachable -information, that is risk, scenario and predicted credit ratings, but also on news and other information sources.*

3.5 Trading

For the last little existence a major topic of research has been the structure of quantitative trading tools using data mining methods based on past data as input to compute short term movements of vital currencies, interest rates, or equities. The goal of this technique is to spot times when markets are cheap or expensive by recognize the factor that are important in formative market returns. The trading system examines the relationship between relevant information and piece of financial assets, and gives you buy or sell recommendations when they suspect an under or overvaluation.



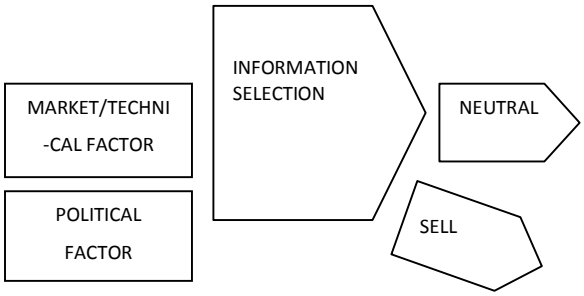


Fig 5: *Market participants examine the relationship between relevant information and the price of financial assets, and buy or sell securities when they suspect an under or over valuation*

4. DATA MINING TECHNIQUES

To determine the main algorithms used for financial accounting fraud detection, we present a Review of data mining techniques is applied to the detection of financial sector.

4.1 Clustering: Clustering is used to division objects into earlier unidentified conceptually meaningful groups (i.e. clusters), with the partitioning and is regarded as a variation of unsupervised classification [8]. Cluster analysis decomposes or partitions a data set (single or multivariate) into dissimilar groups so that the data points in one group are related to each other and are as different as possible from the data points in other groups [6]. It is suggested that data objects in each cluster should have high intra-cluster comparison within the same cluster but should have low inter-cluster resemblance to those in other clusters [7]. The most common clustering techniques are the K-nearest neighbour, the Naïve Bayes technique and self-organizing maps.

4.2 Classification: Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large [10]. Basically classification is used to classify each item in a set of data into one of predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics. In classification, we make the software that can learn

how to classify the data items into groups [9]. Fraud detection and credit risk applications are particularly well suited to this type of analysis. This approach regularly employs decision tree or neural network-based classification algorithms. The data classification procedure involves learning and classification. In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the exactness is adequate the rules can be applied to the new data tuples. For a fraud detection application, this would include complete records of both fake and appropriate actions determined on a record-by-record basis. The algorithm then encodes these parameters into a model called a classifier [6].

4.3 Prediction: Prediction estimates numeric and ordered future values based on the patterns of a data set [11]. It is noted that, for prediction, the attribute, for which the value being predicted is continuous-valued (ordered) rather than categorical (discrete-valued and unordered). This attribute is referred as the predicted attribute [10]. Neural networks and logistic model prediction are the most commonly used prediction techniques.

4.4 Outlier Detection: Outlier detection is employed to measure the distance between data objects to detect those objects that are disgustingly different from or incompatible with the stable data set [8]. Data that appear to have different characteristics than the rest of the population are called outliers [12]. The problem of outlier/anomaly detection is one of the most fundamental issues in data mining. A usually used technique in outlier detection is the discounting learning algorithm [13].

4.5 Regression Models: The regression based models are mostly used in financial accounting fraud detection. The majority of them are based on logistic regression, stepwise-logistic regression, multi criteria decision creation method and exponential pervasive beta two (EGB2) [14]. Logistic model is a complete linear model that is used for binomial regression in which the analyst variables can be also numerical or categorical [17]. It is mainly used to solve problems caused by insurance and commercial fraud. Some of the research has suggested logistic regression based model to predict the existence of financial statement fraud [17]. Statistical method of logistic regression can detect falsified financial statements efficiently [30]. Some researchers have also developed generalized qualitative response model based on Probit and Logit techniques to predict financial statement fraud. That model was based on a dataset collected by an international public accounting company and needs testing for generalization [15]. The study in found that, when the fraud is being executed, insiders, i.e. top executives and managers, reduce their stock holdings through high stock selling activity. The other methods like numerical regression analysis are also useful to test if the existence of a sovereign audit committee mitigates or reduces the likelihood of fraud. The regression analysis using Logit model can be used for observed analysis of financial indexes which can significantly predict financial fraud [16].

Types of regression methods

- Linear Regression
- Multivariate Linear Regression
- Nonlinear Regression
- Multivariate Nonlinear Regression

4.6 Neural Networks: The neural networks are non-linear statistical data modeling tools that are stimulated by the functionality of the human brain using a set of interconnected nodes [20]. Neural networks are widely applied in classification and clustering, and its advantages are as follows. First, it is adaptive; second, it can generate robust models; and third, the classification process can be modified if new training weights are set. Neural networks are chiefly applied to credit card, automobile insurance and corporate fraud. Literature describes that neural networks can be used as a financial fraud detection tool. The neural network fraud classification model

employing endogenous financial data created from the learned behaviour pattern can be applied to a test sample. The neural networks can be used to predict the occurrence of corporate fraud at the administration level. Researchers have explored the effectiveness of neural networks, decision trees and Bayesian belief networks in detecting fraudulent financial statements (FFS) and to identify factors associated with FFS [18]. The study in [19] revealed that input vector consisted of financial ratios and qualitative variables, was more effective when fraud detection model was developed using neural network. The model was also compared with standard statistical methods like linear and quadratic discriminate analysis, as well as logistic regression methods [19]. The generalized adaptive neural network architectures and the adaptive logic network are well received for fraud detection

.

4.7 Bayesian Belief Network: The Bayesian belief network (BBN) represents a set of random variables and their conditional independencies using a directed acyclic graph (DAG), in which nodes represent random variables and misplaced edges encode conditional independencies between the variables [21]. The Bayesian belief network is used in developing models for credit card, automobile insurance, and commercial fraud detection. The research in [21] described that Bayesian belief network model correctly classified 90.3% of the justification sample for fraud detection. Bayesian belief network outperformed neural network and decision tree methods and achieved outstanding classification accuracy [21].

4.8 Decision Trees: A decision tree (DT) is a tree structured decision support tool, where each node represents a test on an attribute and each branch represents possible consequences. In this way, the predictive model attempts to divide annotations into mutually exclusive subgroups and is used for data mining and machine learning tasks [21]. Decision trees are predictive decision support tools that create mapping from explanation to possible consequences [8]. Predictions are represented by leaves and the conjunctions of features by branches. Decision trees are commonly used in credit card, automobile insurance, and corporate fraud. A decision tree is a mapping from observations about an item to conclusion about its target value as a predictive model in data mining and machine learning. Generally, for such tree models, other descriptive names are classification tree (discrete target) or regression tree (continuous target). In these tree structures, the leaf nodes represent classifications, the inner nodes represent the current predictive attributes and branches represent conjunctions of attributions that lead to the final classifications.

4.9 Nearest Neighbour Method: Nearest neighbour method is a similarity based classification approach. Based on a permutation of the classes of the most similar k record(s), every record is classified. Sometimes this method is also known as the k-nearest neighbour technique [8]. K-nearest neighbour method is used in automobile insurance claims fraud detection and for identifying defaults of credit card clients.

4.10 Expert Systems: Researchers in the field of Expert systems have examined the role of Expert Systems in growing the detecting ability of auditors and statement users. By using expert system, they could have better detecting abilities to accounting fraud risk under different context and level and enable auditors give much reliable auditing suggestions through rational auditing procedure. The research has confirmed that the use of an expert system enhanced the auditor's presentation. With assistance from expert system, the auditors discriminated better, among situations with different levels of management fraud-risk. Expert System aided in decision making concerning appropriate audit actions. The main purpose is to apply a hybrid decision support system using stacking variant methodology to detect fraudulent financial statements.

4.11 Genetic Algorithm: Genetic Algorithm (GA) was developed by Holland in 1970. GA is stochastic search algorithm modelled on the process of natural selection, which underlines biological evolution. A has been successfully applied in many search, optimization, and machine learning problems

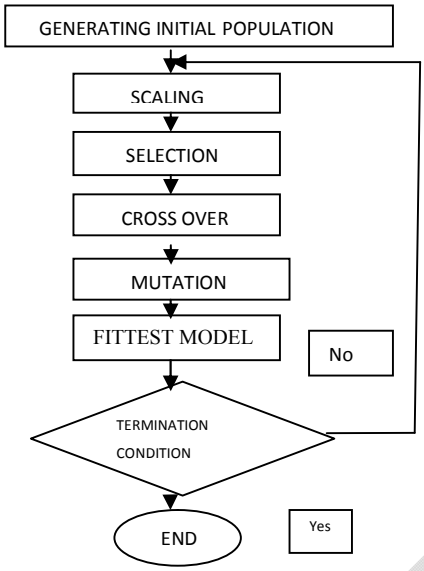


Fig 6: *Genetic algorithm steps*

Every string is the encoded binary, real etc., version of a candidate solution. An estimate function associates a fitness measure to every string representing its fitness for the problem. Standard GA apply genetic operators such selection, crossover and mutation on an originally random population in order to compute a whole generation of new strings

Selection deals with the probabilistic endurance of the fittest, in those more fit chromosomes are chosen to survive. Where fitness is a comparable measure of how well a chromosome solves the problem at hand.

Crossover takes individual chromosomes from P combines them to form new ones.

Mutation alters the new solutions so as to add complicated in the search for better solutions. In general the main motivation for using GAs in the discovery of high-level prediction rules is that they perform a global search and cope better with attribute interface than the greedy rule induction algorithms often used in data mining.

5. Conclusion

This paper aimed to focus the data mining application in the finance sector for extracting useful information. The prediction of risks using Data Mining applications is a challenging task but it drastically reduces the human effort and increases the diagnostic accuracy. The data mining technology may have a positive effect on financial sector. It is concluded that the ability to use data mining is an important criteria for a good financial institution. And also it is found that there is not a meaningful relation between the knowledge of data mining and the intention of accepting data mining technology. Using DM technologies in financial field will render a competitive superiority for the business operations. The business operations which integrate DM techniques with their financial institution, which enables estimation of future based on previous data, enables the interest groups to make pre-emptive estimations and to determine the possible risks.

REFERENCES

[1] Hillol Kargupta, Anupam Joshi, Krishnamoorthy Siva Kumar, Yelena Yesha, "Data Mining: Next Generation Challenges and Future Directions", Publishers: Prentice-Hall of India, Private Limited, 2005.

[2] Muraleedharan.D, "Modern Banking: Theory and Practice",PHI Learning private Limited, 2009

[3] Bharati M. Ramager, "Data Mining Techniques and Applications", International Journal of Computer Science and Engineering.

[4] Desh pande.S.P, Dr. Thakare. V.M, "Data Mining System and Applications: A Review". Risk management

[5] Dr. Madan Lal Bhasin, "Data Mining: A Competitive Tool in the Banking and Retail Industries", the Chartered Accountant October 2006

[6] Yue, X., Wu, Y., Wang, Y. L., & Chu, C. (2007). A review of data mining-based financial fraud detection research, international conference on wireless communications Sep, Networking and Mobile Computing (2007) 5519–5522.

[7] Zhang, D., & Zhou, L. (2004). Discovering golden nuggets: data mining in financial application, IEEE Transactions on Systems, Man and Cybernetics 34 (4) (2004) Nov.

[8] Han, J., & Kamber, M. (2006). Data Mining: Concepts and Techniques, Second edition, Morgan Kaufmann Publishers, 2006, pp. 285–464.
[9] Data Mining Techniques, http://www.dataminingtechniques.net/
[10] Bharati M. Ramageri, "DATA MINING TECHNIQUES AND APPLICATIONS", Indian Journal of Computer Science and Engineering Vol. 1 No. 4
[11] Ahmed, S.R. (2004). Applications of data mining in retail business, International Conference on Information Technology: Coding and Computing 2 (2) (2004) 455–459.
[12] Agyemang, M., Barker, K., & Alhajj, R. (2006). A comprehensive survey of numeric and symbolic outlier mining techniques, Intelligent Data Analysis 10 (6) (2006) 521–538.
[13] Yamanishi, K., Takeuchi, J., Williams, G., & Milne, P. (2004). On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms, Data Mining and Knowledge Discovery 8 (3) (2004) 275–300.
[14] Ngai, E.W.T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2010). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature, Decision Support System (2010), doi:10.1016/j.dss.2010.08.006.
[15] Wang, J., Liao, Y., Tsai, T. & Hung, G. (2006). Technology-based financial frauds in Taiwan: issue and approaches, IEEE Conference on: Systems, Man and Cyberspace Oct (2006) 1120–1124.
[16] Eick, S.G. & Fyock, D.E. (1996). Visualizing corporate data, AT&T Technical Journal 75 (1) (1996) 74–86..
[17] Spathis, C. T. (2002). Detecting false financial statements using published data: some evidence from Greece, Managerial Auditing Journal 17 (4) (2002) 179–191.
[18] Kirkos, E., Spathis, C., & Manolopoulos, Y. (2007). Data mining techniques for the detection of fraudulent financial statements, Expert Systems with Applications 32 (4) (2007) 995–1003.
[19] Fanning, K., & Cogger, K. (1998). Neural network detection of management fraud using published financial data. International Journal of Intelligent Systems in Accounting, Finance & Management, vol. 7, no. 1, pp. 21- 24, 1998.
[20] Silverstone, Howard, & Sheetz, M. (2004). Forensic Accounting and Fraud Investigation for Non-Experts. Hoboken, John Wiley & Sons, 2004.

Authors Biography:

Kumar Charlie Paul, Principal of A.S.L Pauls College of Engineering & Technology. Had did many National and International Conferences and published many papers in journals. He also guided many students for their Ph.D project works. Having more than 23 years of experience in teaching field.

Immanual Gnandurai, Assistant Professor/ CSE of Dhaya college of Engineering. Had did many National and International Conferences and published many papers in journals. He also guided many students for their UG and PG project works.