Data Mining Process Related Problems in Financial Applications

Shaik Mahammad Rafi[1], B.Naveen Kumar[2]

Assistant Professor,CSE Department,AITS,Rajampet,Andhra Pradesh [1,2]

Abstract: ***This paper describes data mining process related problems in financial applications for financial analytics and explores methodologies and techniques in data mining area combined with predictive analytics for application driven results for financial data. The application of DM techniques on financial data can contribute to the solution of classification and prediction problems and facilitate the decision making process The basic idea is to apply patterns on available data and generate new assumptions and anticipated behavior using predictive analysis. It includes applications like Surveillance and Warning Systems, Predicting Abnormal Stock Market Returns, Predicting Corporate Bankruptcies, Financial Distress, Management Fraud. Data mining methods used in these applications are Regression analysis, Choice modeling, Rule Induction, Network/Link Analysis, Clustering /Ensembles, Neural networks, Decision trees, Bayesian data analysis.***

Keywords: Data Mining, Predictive Analytics, Financial Data, Financial Applications, Predicting Corporate Bankruptcies, Financial Distress

I.INTRODUCTION

The Financial data are collected by many organizations like banks, stock exchange authorities, taxation authorities, big accounting and auditor offices specialized data bases, etc and in some cases are publicly available. The application of DM techniques on financial data can contribute to the solution of classification and prediction problems and facilitate the decision making process. Typical examples of financial classification problems are corporate bankruptcy, credit risk estimation, going concern reporting, financial distress and corporate performance prediction. The importance of DM in finance and accounting has been recognized by many organizations.

Specifics of data mining in finance are coming from the need to:

a. Forecast multidimensional time series with high level of noise.

b. Accommodate specific efficiency criteria

c. (e.g., the maximum of trading profit) in addition to prediction accuracy such as R2.

d. Make coordinated multi-resolution forecast (minutes, days, weeks, months, and years).

e. Incorporate a stream of text signals as input data for forecasting models (e.g., Enron case, September 11 and others).

f. Be able to explain the forecast and the forecasting model ("black box" models have limited interest and future for significant investment decisions).

Data Mining (DM) is a well honored field of Computer Science. It emerged in late 80's by using concepts and methods from the fields of Artificial Intelligence, Pattern Recognition, Database Systems and Statistics, DM aims to discover valid, complex and not obvious hidden information from large amounts of data. For this reason, another equivalent term for DM is Knowledge Discovery in Databases (KDD), which is equally often met in the literature. Generally, data mining is the process of analyzing data from different perspectives and summarizing it into useful information. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

Predictive analytics is the branch of the advanced analytics which is used to make predictions about unknown future events. Predictive analytics uses many techniques from data mining, statistics, modeling, machine learning and artificial intelligence to analyze current data to make predictions about future. The patterns found in historical and transactional data can be used to identify risks and opportunities for future. Predictive analytics allows organizations to become proactive, forward looking, anticipating outcomes and behaviors based upon the data and not on a hunch or assumptions. Industry experts are focusing on customers and developing customer-centric projects. Using predictive analytics to accurately anticipate market conditions and
customer behavior enables firms to provide

personalized services that boost customer loyalty and develop new markets. Integrated systems are needed so data can be handled with the constantly growing variety and volume of data.

A financial application is a software program that facilitates the management of business processes that deal with money. Finance is a field that deals with the study of investments. It includes the dynamics of assets and liabilities over time under conditions of different degrees of uncertainty and risk. Finance can also be defined as the science of money management. A key point in finance is the time value of money, which states that purchasing power of one unit of currency can vary over time. Finance aims to price assets based on their risk level and their expected rate of return. Huge electronic data repositories are being maintained by banks and other financial institutions. Valuable bits of information are embedded in these data repositories. The huge size of these data sources make it impossible for a human analyst to come up with interesting information (or patterns) that will help in the decision making process. A number of commercial enterprises have been quick to recognize the value of this concept.

2. Data Mining Process

Data mining, also popularly referred to as knowledge discovery from data (KDD), is the extraction of patterns representing knowledge implicitly stored or captured in large databases, data warehouses, the Web, other massive information repositories or data streams.

The architecture of a typical data mining system may have the following major components

a. Database, data warehouse, or other information repository. This is one or a set of databases, data warehouses, spread sheets, or other kinds of information repositories. Data cleaning and data integration techniques may be performed on the data.

b. Database or data warehouse server. The database or data warehouse server is responsible for fetching the relevant data, based on the user's data mining request.

c. Knowledge base. This is the domain knowledge that is used to guide the search, or evaluate the interestingness of resulting patterns. Such knowledge can include concept hierarchies, used to organize attributes or attribute values into different levels of abstraction. Knowledge such as user beliefs, which can be used to assess a pattern's interestingness based on its unexpectedness, may also be included. Other examples of domain knowledge are additional interestingness constraints or thresholds, and metadata.

d. Data mining engine. This is essential to the data mining system and ideally consists of a set of functional modules for tasks such as characterization, association analysis, classification, evolution and deviation analysis.

e. Pattern evaluation module. This component typically employs interestingness measures and interacts with the data mining modules so as to focus the search towards interesting patterns. It may access interestingness thresholds stored in the knowledge base. Alternatively, the pattern evaluation module may be integrated with the mining module, depending on the implementation of the data mining method used. For efficient data mining, it is highly recommended to push the evaluation of pattern interestingness as deep as possible into the mining process so as to find the search to only the interesting patterns.

f. Graphical user interface. This module communicates between users and the data mining system, allowing the user to interact with the system by specifying a data mining query or task, providing information to help focus the search, and performing exploratory data mining based on the intermediate data mining results. In addition, this component allows the user to browse database and data warehouse schema or data structures, evaluate mined patterns, and visualize the patterns in different forms.

Data mining is iterative. A data mining process continues after a solution is deployed. The lessons learned during the process can trigger new business questions. Changing data can require new models. Subsequent data mining processes benefit from the experiences of previous ones.

2.1 DM Techniques & Predictive Analysis

Predictive analytics, pattern recognition, and classification problems are not new. Long used in the financial services and insurance industries, predictive analytics is about using statistics, data mining, and game theory to analyze current and historical facts in order to make predictions about future events.

Regression analysis : Regression models are the

mainstay of predictive analytics. The linear regression model analyzes the relationship between the response or dependent variable and a set of independent or predictor variables. That relationship is expressed as an equation that predicts the response variable as a linear function of the parameters.

Choice modeling : Choice modeling is an accurate and general-purpose tool for making probabilistic predictions about decision -making behavior. It behooves every organization to target its marketing efforts at customers who have the highest probabilities of purchase. Choice models are used to identify the most important factors driving customer choices. Typically, the choice model enables a firm to compute an individual's likelihood of purchase, or other behavioral response, based on variables that the firm has in its database, such as geo-demographics, past purchase behavior for similar products, attitudes, or psychographics.

Rule induction : Rule induction involves developing formal rules that are extracted from a set of observations. The rules extracted may represent a scientific model of the data or local patterns in the data. One major rule-induction paradigm is the association rule. Association rules are about discovering interesting relationships between variables in large databases. It is a technique applied in data mining and uses rules to discover regularities between products. For example, if someone buys peanut butter and jelly, he or she is likely to buy bread. The idea behind association rules is to understand when a customer does X, he or she will most likely do Y. Understanding those kinds of relationships can help with forecasting sales, promotional pricing, or product placements.

Network/Link Analysis : This is another technique for associating like records. Link analysis is a subset of network analysis. It explores relationships and associations among many objects of different types that are not apparent from isolated pieces of information. It is commonly used for fraud detection and by law enforcement. You may be familiar with link analysis, since several Web-search ranking algorithms use the technique.

Clustering/Ensembles : Cluster analysis, or clustering, is a way to categorize a collection of "objects," such as survey respondents, into groups or clusters to look for patterns. Ensemble analysis is a newer approach that leverages multiple cluster solutions (an ensemble of potential solutions). There are various ways to cluster or create ensembles. Regardless of the method, the purpose is generally the same—to use cluster analysis to partition into a group of segments and target markets to better understand and predict the behaviors and preferences of the segments. Clustering is a valuable predictive-analytics approach when it comes to product positioning, new-product development, usage habits, product requirements, and selecting test markets.

Neural networks : Neural networks were designed to mimic how the brain learns and analyzes information. Organizations develop and apply artificial neural networks to predictive analytics in order to create a single framework. The idea is that a neural network is much more efficient and accurate in circumstances where complex predictive analytics is required, because neural networks comprise a series of interconnected calculating nodes that are designed to map a set of inputs into one or more output signals. Neural networks are ideal for deriving meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by humans or other computer techniques. Marketing organizations find neural networks useful for predicting customer demand and customer segmentation.

Decision trees : Decision trees use real data-mining algorithms to help with classification. A decision-tree process will generate the rules followed in a process. Decision trees are useful for helping you choose among several courses of action and enable you to
explore the possible outcomes for various options in order to assess the risk and rewards for each potential course of action. Such an analysis is useful when you need to choose among different strategies or investment opportunities, and especially when you have limited resources.

Memory-based reasoning (MBR)/Case-based reasoning : This technique has results similar to a neural network's but goes about it differently. MBR looks for "neighbor" kind of data rather than patterns. It solves new problems based on the solutions of similar past problems. MBR is an empirical classification method and operates by comparing new unclassified records with known examples and patterns.

Bayesian classification : Bayesian approaches are a fundamentally important DM technique. Given the

probability distribution, Bayes classifier can provably achieve the optimal result. Bayesian method is based on the probability theory. Bayes Rule is applied here to calculate the posterior from the prior and the likelihood, because the later two is generally easier to be calculated from a probability model.

Uplift modeling : It is also known as incremental-response modeling. This technique directly models the incremental impact of targeting marketing activities. The uplift of a marketing campaign is usually defined as the difference in response rates between a treated group and a randomized control group. Uplift modeling uses a randomized scientific control to measure the effectiveness of a marketing action and to build a model that predicts the incremental response to the marketing action. Predictive modeling mathematically represents underlying relationships in historical data in order to explain the data and make predictions, forecasts or classifications about future events. Predictive models summarize large quantities of data to amplify its value. Predictive models improve marketing effectiveness by analyzing past performance to assess how likely a customer is to exhibit a specific behavior in the future.

3. Financial Applications with Data Mining With Predictive Analysis

Continuous predictive analysis means you can extrapolate what has happened so far to predict that something might be about to happen and prevent it. Consider a wild algorithm. Under normal circumstances you can be monitoring the algorithm's operating parameters, which might include what instruments are traded, size and frequency of orders, order-to-trade ratio etc. Real-time monitoring means actions can be taken in time to have an impact on the business. Using customer data, banks and other financial institutions are applying the technology to predict customers likely to churn and then taking action to prevent the churn from occurring. Predictive analytics identifies customers likely to churn, then segments those customers by profitability, volume, and length of engagement.

3.1 Surveillance and Warning Systems : Market surveillance today means watching everyone and everything at once. It means sniffing out abnormal trader behavior while, at the same time, monitoring markets for possible manipulation -- and reading news headlines while checking on chat rooms for possible wrongdoing. These systems are basically designed for collecting huge market data and identify market manipulations, insider trading, fraud and compliance violations and also include ensuring rule compliance.

The three activity streams are

i) Collect and Create
ii) Detect & Investigate
iii) Enforce & Discipline

One of the best examples of predictive analytics is seen at Financial Industry Regulatory Authority. FINRA is an independent regulator overseen by the Securities and Exchange Commission, chartered to monitor the U.S. stock market. FINRA oversees every brokerage firm and broker doing business with the U.S. public and monitors trading on the U.S. stock markets.

3.2 Predicting Abnormal Stock Market Returns :

Insider traders usually make abnormal returns because of the insider information available. Outsiders who can get access to the insider information can also make increased profits. The ability of outsiders, using insider trading information, to predict abnormal returns can be increased by focusing on data such as the size of the company and the number of months in the future that are predictive for stock prices. In this subject, the studies conducted by Safer, (2002) may be summarized as follows.

The insider trading data used in this study are from January 1993 to mid June 1997.The data was collected from Securities and Exchange Commission. The stocks used in the analyses included all stocks in the S&P 600 (small cap), S&P 400 (midsize cap) and S&P 500 (large cap) as of June 1997 that had insider records for the entire period of the study. There were 946 stocks in the three market caps which had available data in January1993. From the list of 946 stocks, the sample included every stock that averaged at least two buys per year. This resulted 343 stocks being used for the study. The variables in the original data set include the company, name and rank of the insider, transaction date, stock price, number of

shares traded, type of transaction (buy or sell), and number of shares held after the trade. To assess an insider's prior trading patterns, the study examined the previous nine and 18 weeks of trading history. The prediction time frames for predicting abnormal returns were established as three, six, nine, and 12 months. Then the data can be split into a training set (80% of the data) and validation set (20%). A neural network model is applied. Safer found that the prediction of abnormal returns could be enhanced in the following ways:

• Extending the time of the future forecast up to one year.
• Increasing the period of back aggregated data;
• Narrowing the assessment to certain industries such as electronic equipment and business services and
• Focusing on small and midsize rather than large companies.

3.3 Predicting Corporate Bankruptcies : Bankruptcy prediction seems to be most popular topic of the application of DM techniques on financial data. Corporate bankruptcy causes economic damages for management, investors, creditors and employees together along with social cost. For these reasons bankruptcy prediction is an important issue in finance. Bankruptcy prediction by using financial statements data attracts its origin from the work of Altman in 1968. Altman argues that corporate failure is a long period process and that the financial statement data should include warning signals for the imminent bankruptcy. By applying Multiple Discriminant Analysis techniques he developed a model for bankruptcy prediction. Since the work of Altman many researchers developed alternative models by using statistical techniques (Ohlson 1980 used Logit, Zmijewski 1984 used Probit). In the last years research effort has been made to build models which use DM techniques. Lin and McClean (2001) tried to predict corporate failure by using four different methods. Two of the methods are statistical (Discriminant Analysis and Logistic Regression), whereas the remaining two methods are Machine Learning techniques (Decision Trees – C5.0 and Neural Networks). The model was applied to predict bank failures. Input variables are nine financial variables, which have been found to be significant in previous studies. The sample contained data about 2555 non failed and 548 failed banks. 20% of the data were used as training set and 80% as testing set.

To reduce Type one errors the sample was balanced to include equal number of failed and not failed banks. Two feature selection methods have been employed reducing the input variables to 4 by using human judgment and to 15 by using ANOVA. The authors report better results for the NNs and decision trees models for both the human judgment based and the ANOVA feature selection. Finally, the authors propose a hybrid algorithm employing weighted voting of different classifiers. Marginally better performance is reported for the hybrid model.

3.4 Financial Distress : According to SAS 59, the auditor has to evaluate the ability of his/her client to continue as a GC for at least one year beyond the balance sheet data. If there are indications that the client company will face financial difficulties, which may lead to failure, the auditor has to issue a going concern report. The assessment of the going concern status is not an easy task. Studies report that only a relative small proportion of failed firms have been qualified on a going concern basis (Koh 2004). To facilitate the auditors on the going concern report issuing task, statistical and machine learning techniques have been proposed. Koh (2004) compared back propagation NN, Decision Trees and logistic regression methods in a going concern prediction study. The data sample contained 165 going concern firms and 165 matched non going concern firms ,selected financial ratios have been used as input variables. The author reported that Decision Trees outperformed the other two methods

Tan and Dihardjo (2001) built upon a previous study of Tan, which tried to predict financial distress for Australian credit unions by using NNs. In his previous study Tan used quarterly financial data and tried to predict distress in a quarter base. Tan and Dihardjo improved the method by introducing the notion of "early detector". When the model predicts that a credit union will go distressed in a particular quarter and the union actually goes distressed in a next quarter, in a maximum of four quarters, the quarter is labeled as "Early Detector". This improved method performed better than the previous one in terms of Type II errors rate. 13 financial ratios were used as input variables and a sample of 2144 observations was used. The results were compared with those of a Probit model and were found marginally better especially for the Type 1

error rate. Konno and Kobayashi (2000) proposed a method for enterprise rating by using Mathematical Programming techniques. The method made no distribution assumptions about the data. Three alternatives based on discrimination by hyperplane, discrimination by quadratic surface and discrimination by elliptic surface were employed. 6 financial ratios derived from financial statements were used as input variables. The method calculated a score for each enterprise.

3.5 Management Fraud : Management fraud is the deliberated fraud committed by managers through falsified financial statements. Management fraud injures tax authorities, share holders and creditors. Spathis (2002) developed two models for identifying falsified financial statement from publicly available data. Input variables for the first model contain 9 financial ratios. For the second model z-score is added as input variable to accommodate the relationship between financial distress and financial statement manipulation. The method used is logistic regression and the data sample contained 38 FFS and 38 non FFS firms. For both models the results show that three variables with significant coefficients entered the model. Researchers in the field of Expert systems have examined the role of Expert Systems in increasing 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 auditors" performance. With assistance from expert system, the auditors discriminated better, among situations with different levels of management fraud-risk. Expert System aided in decision making regarding appropriate audit actions. The financial accounting fraud detection research is classified as per data mining application and data mining techniques. Some researchers have tried to apply a combination of many data mining techniques like decision trees, neural networks, Bayesian belief network, K-nearest neighbor. The main objective is to apply a hybrid decision support system using stacking variant methodology to detect fraudulent financial statements.

The research related with application of data mining algorithms and techniques for financial accounting fraud detection is a well studied area. The implementation of these techniques follows the same information flow of data mining processes in general. The process starts with feature selection then proceeds with representation, data collection and management, pre-processing, data mining, post-processing, and in the end performance evaluation.

4. Conclusions

Next future direction is developing practical decision support software tools that make easier to operate in data mining environment specific for financial tasks, where hundreds and thousands of models such as neural networks, and decision trees need to be analyzed and adjusted every day with a new data stream coming every minute and monitoring the stock market. Inside of the field of data mining in finance we expect an extensive growth of hybrid methods that combine different models and provide a better performance than can be achieved by individuals. In such integrative approach individual models are interpreted as trained artificial "experts". Therefore their combinations can be organized similar to a consultation of real human experts. Moreover, these artificial experts can be effectively combined with real experts. It is expected that these artificial experts will be built as autonomous intelligent software agents. Thus "experts" to be combined can be data mining models, real financial experts, trader and virtual experts that runs trading rules extracted from real experts. A virtual expert is a software intelligent agent that is in essence an expert system. We coined a new term "expert mining" as an umbrella term for extracting knowledge from real human experts that is needed to populate virtual experts. We also expect that the blending with ideas from the theory of dynamic systems, chaos theory, and physics of finance will deepen.

Predictive analytics software is increasingly easier to use, it's no surprise the technology is being adopted more and more in the financial services industry. Using customer data, banks and other financial institutions are applying the technology to predict customers likely to churn and then taking action to prevent the churn from occurring. Predictive analytics identifies customers likely to churn, then segments those customers by profitability, volume, and length of engagement. Once segmented, banking business analysts, often working in tandem with marketing and sales teams, again apply the technology to optimize marketing campaigns that ensure exactly the correct incentives are offered to each class of customers. This

results in higher retention rates at lower costs, and can improve the customer experience by more precisely offering promotions that appeal to them. Financial services institutions also use predictive analytics to segment customers and predict which ones will react well to cross-selling promotions. Predictive analytics is ideal for classifying which customers are likely to respond to offers for additional products and services, allowing banks to achieve profitability in the near term, and add to the bottom line over time.

References

[1] Azoff, E., Neural networks time series forecasting of financial markets, Wiley,1994.

[2] Wang J., Data Mining; opportunities and challenges, Idea Group, London, 2003

[3] K.S. Shin and Y.J. Lee: "A Genetic Algorithm Application in Bankruptcy Prediction Modeling", Expert Systems with Applications, Volume 23, Issue 3, October, 2002, pp.321-328.

[4] C. Spathis: "Detecting False Financial Statements Using Published Data: some Evidence from Greece", Managerial Auditing Journal, Volume 17,No 4, 2002, pp.179-191.

[5] Jiawei Han, Micheline Kamber, "Data Mining Concepts and Technique", 3rd edition

[6] Michal Meltzer, Using Data Mining on the road to be successful part III, published in October 2004, retrieved 2nd January

[7] Insurance Information Warehouse (IIW) General Information Manual Transforming Insurance Information into Business Intelligence

[8] Kovalerchuk B, Vityaev E (2000). Data Mining in Finance: Advances in Relational and Hybrid Methods, Kluwer.

[9] Lee LW, Wang LW, Chen SM, Leu YH (2006). Handling forecasting problems based on two-factor high-order time series, IEEE Transactions on Fuzzy Systems.

[10] Tahseen AJ, Syed MAB (2008). A refined fuzzy time series model for stock market forecasting, PHYSICA A, Physica.

Authors

Mr. Shaik.Mahammad Rafi. He was born in Rajampet, Kadapa, A.P, India in 1990.He is working as Asst.Professor in the department of Computer Science and Engineering at Annamacharya Institute of Technology and Sciences,Rjampet,Kadapa,Andhra Pradesh. He has done Bachelor's of Technology from JNTUA University in the year 2011 in Information Technology. He has done his Master of Technology from JNTUA University in the year 2013 in Global college of Engineering ,Kadapa,Andhra Pradesh. Her research areas include Data Mining, Financial Applications and Investments.

Ir. Basetty Naveen Kumar. He Kadapa, A.P, India in 1987.He ofessor in the department of and Engineering at ndhra Pradesh. He has done gy from JNTUH University in ter Science & Engineering. He of Technology from JNTUA r 2011 in SVCET,Chittoor research areas include Data cations and Investments.