ICCES, vol.11, no.2, pp.39-46

Risk modeling by CHAID decision tree algorithm

A.S. Koyuncugil[1], N. Ozgulbas[2]

Summary

In this paper, a data mining model for detecting financial and operational risk indicators by CHAID Decision Tree is presenting. The identification of the risk factors by clarifying the relationship between the variables defines the discovery of knowledge from the financial and operational variables. Automatic and estimation oriented information discovery process coincides the definition of data mining. During the formation of model; an easy to understand, easy to interpret and easy to apply utilitarian model that is far from the requirement of theoretical background is targeted by the discovery of the implicit relationships between the data and the identification of effect level of every factor.

keywords: Data Mining; CHAID Decision Tree Algorithm; Risk Management; Financial Risk; Operational Risk.

Introduction

The effective management and identification of risk is a complicated task, as well as a fundamental element of business success or failure. Nowadays, Basel II Capital Accord, which will become effective in 2009, has been the center of attention of both credit companies and firms. Basel II is the second of the 'Basel Accords', which are recommendations on banking laws and regulations issued by the 'Basel Committee on Banking Supervision'. The purpose of Basel II, which was initially published in June 2004, is to create an international standard that banking regulators can use when creating regulations about how much capital banks need to put aside to guard against the types of risks banks face. Basel-II, which introduces risk-based capital management and risk-based credit pricing, would negatively/positively affect amount/price of loans to be utilized by firms. Under Basel-II banks will from now on not only consider financial risk of firms but also operational risk thereof before they grant loans to firms. With minimum financial and operational risk firms would get higher ratings from banks and independent auditing institutions thereby increasing their chances to receive loan facilities with more favorable conditions and minimum costs (Bank for International Settlements, 2008).

The new capital requirements for lending to firms could be too high especially for SMEs. The majority of the transition economies have acknowledged that SMEs are an engine of economic growth and a source of sustainable development, crucial for industrial restructuring and for the generation of new jobs, creating income for the population; although unemployment is still an increasingly problem among

[1] Research Department Capital Markets Board of Turkey Ankara, Turkey

[2] Department of Healthcare Management Baskent University Ankara, Turkey

these countries. The value of the asset correlation for that type of firm was substantially lower than that for larger firms. An excessive level of capital might have discouraging effects on the willingness of banks to grant loans to SMEs. In particular, a substantial increase in the wedge between regulatory capital needs for SMEs versus those for larger firms might result in a shortage of finance for the former. SMEs are expected to experience problems in receiving an appropriate credit rating and in obtaining low-cost credits from the banks, unless they take the necessary steps to improve their corporate structure and use effective risk management techniques. Benefits of risk management can summarize as early warning to avoid distress, road maps for good credit rating, better business decision making, and greater likelihood of achieving business plan and objectives (OECD,2004; TBAK, 2004; EU, 2008; Bank for International Settlements, 2008).

Most of studies and reports on SMEs in Turkey emphasize that insufficiencies of capital and lack of financial management are the most important problems like Europe and other countries (EU, 2008; BRSA, 2005). Turkish SMEs have to take risks to grow but manage risks to protect the business. These findings and trends in sector indicated that SMEs need risk management to deal with not only Basel-II, but also increasing globalization, negotiations towards full accession to the EU, financial and economic regulations introduced in the markets where SMEs are operating.

This paper presents a data mining risk model for detecting financial and operational risk indicators by CHAID Decision Tree Algorithm. This study was designed as a project to bring out the financial and operational risk factors into open as early warning signals for SMEs in Turkey. Our SME project is the biggest study as covered amount and also first study that designed a data mining model for financial and operation risk in Turkey. The Scientific and Technological Research Council of Turkey has been funded this project.

Model

The identification of the risk factors by clarifying the relationship between the variables defines the discovery of knowledge from the financial and operational variables. Automatic and estimation oriented information discovery process coincides the definition of data mining. During the formation of model; an easy to understand, easy to interpret and easy to apply utilitarian model that is far from the requirement of theoretical background is targeted by the discovery of the implicit relationships between the data and the identification of effect level of every factor.

The risk model has five phases. Data for model collected in the first and second phases of study and these phases completed. We have studied on the other three phases. Completed and other phases of the study and methodologies are presented below.

The first phase of study consists of the process of determining the financial variables of SMEs which will be used in data mining model. Such data obtained by means of financial analyses of balance sheets and income statements of companies available through Turkish Central Bank. Available data of all firms collected between the years of 1994-2006.

In this phase firstly, firms classified in accordance with EU's SME criteria on basis of amount of annual turnover. When classifying the firm's annual turnovers thereof will convert into Euro at average FX rate of Central Bank of the Republic of Turkey (CBRT) and firms were classified into four categories as follows:

- Micro firms,
- Small enterprises,
- Medium enterprises, and
- Large-scale firms (CBRT, 2008).

After the classification firms conforming to the SME definition of EU, financial ratios calculated as a financial risk factors. Also, financial performance level of SMEs determined by ratio analysis. Table 1 shows the ratios and definitions that will be use in the first step of model.

Operational data which couldn't be access by balance sheets and income statements such financial management requirements of SMEs, training and skills of employees and managers, process and systems in SMEs collected in phase 2. To collect data information to be used as operational risk factors data in model, a questionnaire designed and information collected from SMEs via a field study.

In the third phase, qualitative and quantitative data to be obtained through phases 1 and 2 will be analyzed with data mining. The main approach for model is discovering different financial and operational risk factors, levels and early warning signs. Therefore, the model should focus segmentation methods. In the scope of the methods of data mining,

- Logistic regression,
- Discriminant analysis,
- Cluster analysis,
- Hierarchical cluster analysis,
- Self Organizing Maps (SOM),
- Classification and Regression Trees (C&RT),
- CHi-Square Automatic Interaction Detector (CHAID)

can be the principal methods, in addition to this several classification/segmentation methods can be mentioned (Kovalerchuk, Vityaev, 2000; Breiman, Freidman, Olshen, Stone, 2005; Thearling, 2008; Koyuncugil, 2006; Koyuncugil, Ozgulbas, 2008). However, during the development of model for SMEs, one of the basic objectives is to help SME administrators and decision makers, who does not have financial expertise, knowledge of data mining and analytic perspective, to reach easy to understand, easy to interpret, and easy to apply results about the risk condition of their enterprises. Therefore, decision tree algorithms that are one of the segmentation methods can be used because of their easy to understand and easy to apply visualization.

Table 1: Variables and definitions

Ratios	**Definition**
Return on Equity	Net Income / Total Assets
Return on Assets	Net Income/ Total Equity
Profit Margin	Net Income/ Total Margin
Equity Turnover Rate	Net Revenues / Equity
Total Assets Turnover Rate	Net Revenues / Total Assets
Inventories Turnover Rate	Net Revenues / Average Inventories
Fixed Assets Turnover Rate	Net Revenues / Fixed Assets
Tangible Assets to Long Term Liabilities	Tangible Assets / Long Term Liabilities
Days in Accounts Receivables	Net Accounts Receivable/ (Net Revenues /365)
Current Assets Turnover Rate	Net Revenues/ Current Assets
Fixed Assets to Long Term Liabilities	Fixed Assets / Long Term Liabilities
Tangible Assets to Equities	Tangible Assets /Equities
Long Term Liabilities to Constant Capital	Long Term Liabilities / Constant Capital
Long Term Liabilities to Total Liabilities	Long Term Liabilities / Total Liabilities
Current Liabilities to Total Liabilities	Current Liabilities / Total Liabilities
Total Debt to Equities	Total Debt / Equities
Equities to Total Assets	Total Equity/Total Assets
Debt Ratio	Total Dept/Total Assets
Current Account Receivables to Total Assets	Current Account Receivables/ Total Assets
Inventories to Current Assets	Total Inventories / Current Assets
Absolute Liquidity	(Cash+Banks+ Marketable Sec.+ Acc. Rec.) / Current Liab.
Quick Ratio (Liquidity Ratio)	(Cash+Marketable Sec.+ Acc. Rec.)/ Current Liab.
Current Ratio	Current Assets/ Current Liabilities

Decision tree algorithms were suitable for profiling because they are visual and easy-to-understand, easily interpretable, and they allow establishment of rules. With the series of rules obtained from decision trees would be possible to create profiles of firms and then classify firms in terms of levels of financial distress by using such profiles. Therefore, the most important risk indicators of financial distress signals as an early warning can be determined for each profile.

There are different decision tree algorithms. In the late 1970s J. Ross Quinlan introduced a decision tree algorithm named ID3. ID3 picks predictors and their splitting values based on the gain in information that the split or splits provide. ID3 was later enhanced in the version called C4.5. Classification and Regression Trees or CART, a relatively new and popular non-parametric analysis technique, was used after these algorithms. Another equally popular decision tree technology to CART is CHAID or Chi-Square Automatic Interaction Detector. CHAID is similar to CART in that it builds a decision tree but it differs in the way that it chooses its splits. Instead of the entropy or Gini metrics for choosing optimal splits the technique relies on the chi square test used in contingency tables to determine which categorical predictor is furthest from independence with the prediction values (Kovalerchuk, Vityaev, 2000; Breiman, Freidman, Olshen, Stone, 2005).

One of the most important differences between CHAID and the other methods is tree generating. ID3. C 4.5 and CART generate binary trees, whereas CHAID can generate nonbinary trees. CHAID works with all types of continuous or categorical variables. However, continuous predictor variables automatically categorized for the purpose of the analysis. By means of Chi-Square metrics CHAID is able to separately segment the groups classified in terms of level of relations. Therefore, leaves of the tree have not binary branches but as much branches as the number of different variables in the data (Berson et al., 2000). Hence; the method of CHAID is used in the scope of this study.

In phase 4, the fitness and availability of the model for necessities of SMEs will be tested. One of the most important reasons of financing problem that SMEs encounter in Turkey is shortcomings in financial management. Therefore this phase has a vital importance to design model in a manner suitable for use of SMEs' managers.

In the last phase, design of the model will be revised according to the findings of the phase 4. After revising the model will be finalized. Our model can be used to detect financial and operational risk indicators of SMEs. It also gives the early warning signs for financial distress.

Implementation

Implementation of the study is realized according to the data flow diagram of the risk model which is given below in Fig. 1. Implementation is given below by phases.

In Phase 1 predefined indicators was computed from database of Central Bank of The Republic of Turkey (CBRT). CBRT database is included financial data of 143.594 companies belong to years 1992-2006 which is shown in Table 2 below. Table 2 is showed the distribution of companies due to their sizes according to the EU standards. There are 96.179 large scale, 29.829 medium sized, 10.319 small sized, 7.267 micro sized companies' financial data belong to years 1992-2006 years in CBRT database. Predefined financial indicators were computed in sectoral level.

In Phase 2 operational data which couldn't be access by balance sheets and income statements such financial management requirements of SMEs, training and skills of employees and managers, process and systems et. To collect data to be used as operational risk factors data in model, a questionnaire designed and infor-

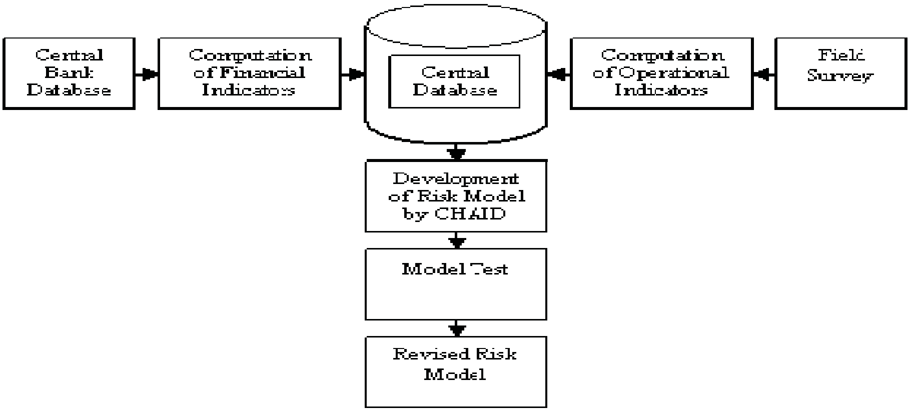
Central Bank Database
Computation of Financial Indicators
Central Database
Computation of Operational Indicators
Field Survey
Development of Risk Model by CHAID
Model Test
Revised Risk Model

Figure 1: Data flow diagram of

mation collected from SMEs via a field study. Field survey was conducted as a Pilot study and it was implemented in OSTIM Organized Industrial Zone in Ankara-Turkey. There were 6.185 companies in OSTIM and the field survey was designed in complete coverage. Therefore, 6.185 companies were taken into consideration in field survey and 6.110 companies were gave response the questionnaire with face to face interviews. Response rate for the field survey 98.8 % and field survey was statistically significant for the representation of OSTIM.

Table 2: Number of Componies in CBRT Database

Year	**Number of The Companies**					**Proportion (%)**
	Large Scale	**Medium Sized**	**Small Sized**	**Micro Sized**	**Total**	
1992	1.369	2.263	1.240	498	5.370	3,74
1993	1.747	2.682	1.653	716	6.798	4,73
1994	4.521	3.536	1.678	1.145	10.880	7,58
1995	9.217	3.405	1.076	873	14.571	10,15
1996	9.034	2.647	774	639	13.094	9,12
1997	7.788	1.527	356	461	10.132	7,06
1998	6.944	1.040	232	298	8.514	5,93
1999	6.129	1.826	537	348	8.840	6,16
2000	6.842	1.751	430	383	9.406	6,55
2001	6.640	1.741	479	374	9.234	6,43
2002	7.400	2.154	580	439	10.573	7,36
2003	7.134	1.800	462	352	9.748	6,79
2004	7.494	1.453	347	311	9.605	6,69
2005	7.430	1.218	271	245	9.164	6,38
2006	6.490	786	204	185	7.665	5,34
Total	96.179	29.829	10.319	7.267	143.594	100,00
Proportion (%)	66,98	20,77	7,19	5,06	100,00	

In Phase 3 financial and operational data will be combined together in central database. Financial indicators will be added to operational records in sectoral level. Therefore, all company records will have both financial and operational variables together. CHAID Decision Tree Algorithm will be used on that combined data for development of risk model. Risk model will be identified financial and operational risk indicators according to the importance level with statistical significancy.

In Phase 4 availability of the model will be tested in selected companies. Companies will be selected via Stratified Random Sampling (SRS) method with 95 % statistical significancy level.

In Phase 5 the risk model will be revised according to the results of Phase 4. The revised risk model will taken into consideration as the final risk model. Therefore, financial and operational risk indicators will be identified from this revised risk model.

As mentioned in part 2 above Phase 1 and Phase 2 was completed and Phase 3 is still on going. After completion of all Phases it is planned that the risk model defined will be turned into software for SMEs.

Acknowledgement

This paper is based on a project which was funded by The Scientific and Technological Research Council of Turkey (TUBITAK).

References

1. Bank for International Settlements. (2008): Basel II: Revised international capital framework. `http://www.bis.org/publ/bcbsca.htm`.

2. Banking Regulation and Supervision Agency (BRSA). (2005): The Possible Effects of Basel II on SMEs Loans. Ankara.

3. Berson, A., Smith, S., and Thearling, K. (2000): Building Data Mining Applications for CRM. McGraw-Hill, USA.

4. Breiman, L., Freidman, J. H., Olshen R. A. and Stone, C. J. (1984): Classification and Regression Trees. Wadsworth.

5. Central Bank of the Republic of Turkey (CBRT). (2008): www.tcmb.gov.tr.

6. EU. Analysis of Competitiveness. (2008): `http://ec.europa.eu/enterprise/enterprise_policy/analysis/observatory_en.htm`.

7. Koyuncugil, A. S. (2006): Fuzzy Data Mining and its application to capital markets. Ph.D. dissertation, Dept. Statistics, Ankara University, Ankara, Turkey,

8. Koyuncugil. A. S. and Ozgulbas, N. (2008): Early Warning System for SMEs as a financial risk detector in Data Mining Applications for Empowering Knowledge Societies. Hakikur Rahman, Ed, Idea Group Inc., USA.

9. Kovalerchuk, B. and Vityaev, E. (2000): Data Mining in Finance. USA: Kluwer Academic Publisher, Hingham MA.
10. OECD. (2004): Small and Medium-Sized Enterprises In Turkey Issues And-Policies Organization For Economic Co-Operation And Development, OECD Press. 2004.
11. The Banks Association of Turkey (TBAK). (2004): Risk Management and the Effect of Basel II on SMEs. Ankara: Publication of TBAK.
12. Thearling, K.(2005): www.thearling.com.
13. Turkish Statistic Institute (TSI). (2006): General Industrial Enterprise Census. http://www.die.gov.tr/TURKISH/SONIST/GSIS/gsisII141003.pdf.