Journal of Optimization in Industrial Engineering 9 (2011) 37-45

Modelling Customer Attraction Prediction in Customer Relation Management using Decision Tree: A Data Mining Approach

Abolfazl Kazemi[a,*], Mohammad Esmaeil Babaei[b]

[a] *Assistant Professor, Department of management, Qazvin Branch, Islamic Azad University, Qazvin, Iran*

[b] *MSc, Department of management, Qazvin Branch, Islamic Azad University, Qazvin, Iran*

Received 24 December, 2011; Revised 16 February, 2011; Accepted 14 March, 2011

Abstract

In Today's quality- based competitive world, known as knowledge age, customer attraction is of ultimate importance. In respect to the slogan "customer is always right", customer relation management is the core of an organizational strategy playing an important role in four aspects of customer identification, customer attraction, customer retaining, and customer satisfaction. Commercial organizations have perceived increased value of customers through analysis of customers' life cycle. Data storing and data mining tools along with other customer relation management methods have provided new opportunities for the business. This paper tries to help organizations determine the criteria needed for the identification of potential customers in the competitive environment of their business by employing data mining in practice. It also provides a mechanism for the identification of potential customers liable to becoming real customers. Using Decision Tree tool, the main criteria are identified and their importance are determine din this paper and then assuming that each main criterion consists of several sub-criteria, their importance in turning potential customers into real ones is in turn determined. By utilizing the identified criteria and sub- criteria, organizations are able to drive selling processes in each attendance in a direction which results in attendants' (future customers) purchase.

Keywords: Data mining (DM); Decision tree (DT); Customer relationship management (CRM).

1. Introduction

Nowadays, manufacturing organizations, considering the role of production in a competitive market system, find themselves in a fully changed environment. In the business models of the past decades with traditional production method approach, supply was less than demand, there existed producer-superiority, the final aim of producers was to produce as much as possible with no intention to improve quality, products were less varied, cost was the most important competitive factor, and products became outdated after a very long term. Development of integrated production management systems and use of new methods in management, selection of appropriate production strategies, and progress in electronic systems by development of information and communication technologies changed the industrial world drastically. In today's business models, using new production methods and considering costumers as the sole focus of production companies and with the purpose of obtaining their maximum satisfaction, supply has exceeded demand. The producer-superiority is replaced with customer-superiority. In addition, it is the customer who determines which products to be produced. Increased production is no longer the main goal, rather it is quality which counts. Product life cycle has become short and products are highly varied. Accountability is the most important competitive factor. Today, organizations are pushed to offer higher quality products and services.

Customer attraction is highly important today. In competitive business environments, the ability to identify profitable costumers and consistent development of existing relationships are the key competitive factors of an organization. Customer Relation Management (CRM) helps organizations recognize customers' value, identify the most profitable ones and develop high quality relations with them. Careful examination of profitability of each customer and identification of the most profitable ones are key factors of CRM success (Cheng et al. 2009).

CRM is a key commercial strategy in which an organization focuses on its customer's needs and should develop a customer-oriented approach in the entire organization (Cheng et al. 2009). A perfect commercial process includes customer attraction, customer increase, and customer retention (Yong et al. 2008). Data mining and knowledge discovery in databases were formed with

* Corresponding author E-mail: abkaazemi@gmail.com

the emergence and use of databases in the early 1980s. One of the most important tasks of an organization is to identify customers and invest on the potential ones. If an organization could easily know which potential customer woul turn into real customers, it could easily identify, attract, retain and develop them (Rygielski et al. 2002).

Data extraction is a new method for organizations to identify customer trends and improve their relations with customers. It is one of the most important tools known in CRM.

Using data mining models, lots of studies have been carried out in CRM, however there has been no model developed by which the identification of effective variables in turning potentials into real customers in the customer attraction step, could in turn increase the actualization of selling process (purchase done by the attendant). The models presented are mainly related to the introduction of solutions after customers' receiving services from organization. In the model presented in this paper, identification of effective factors of sell increase is focused in an attempt to attract attendants and offer services to them, retain and develop them. In so doing, it tries to turn attendants into loyal customers and increase organization investment on the group of customers who purchase according to the model. This model cuts the organization's expenses used for market development and increasing shares of the market.

2. Literature Review

Lowell (1983) could be considered as the first scholar who presented a report on data mining called "simulation of data mining operation". At the same time, researchers and experts of computer, statistics, artificial intelligence, machine learning, etc. began researching in the same area and the related fields. Serious research on data mining started from early 1990s. Many researches and seminars were carried out, training courses, and conferences were also held (sinceStefano et al. 2004).

Today, some data extraction methods such as decision tree (DT), artificial neural network(ANN), genetic algorithm (GA),association rules (AR) which are primarily used in engineering, science, financial and business, are employed in CRM in order to solve the problems concerning customers. Decision tree is a structure like a hanging diagram in which each internal stem shows an experiment (test) and each branch shows test output and the leaves present degree or degrees dispersion. Artificial neural network includes a series of connected processing structures (neurons) which could process information and show advanced input and output relations using statistical model, computational model or non- linear statistical data model tools. Genetic algorithm, presented in 1970 by John Holland in Michigan University, is a search algorithm for solving computer problems based on regular selection and evolution process. Lee and Park (2005) described RS theory and RFM model in their research and suggested different advantages of these two methods in classification of customers in a case study.

In 2004, Baesens et.al, in addition to pointing to K.D.D concept, presented several fuzzy clustering mathematical techniques based on fuzzy algorithms. He showed that fuzzy methods are helpful in clustering. His work focuses on knowledge discovery and extraction from large databases.

At the same time, Hwang et al. while pointing out the importance of the attraction and retention of profitable customers for organizations, believed that calculation of customer value is very difficult indeed. Using LTV, they presented a framework to analyze customer value. They divided customer value into three categories:"current value, potential value, and loyalty value", and only based on these three categories could they calculate customer value. They provided in-depth definition and comprehensive literature of LTV models and also presented a new one (Hwang et al. 2004).

In 2008, Nagi et.al classified available articles between 2000 and 2006considering four aspects of customer relation management (costumer identification, customer attraction, customer retention, customer development) and data mining aspects (classification, clustering, prediction, regression, etc.).

At the same time, Cheng and Chen (2009) presented a model for classification of customer value based on RFM characteristics and K-Means algorithm; RFM model is considered as an input characteristic. Then they calculated the quantitative value of K-Means classification. A four-stage innovative method was also presented.

In another research in 2008, Korach and Stern studied the literature of decision tree method, clustering and general issues related to Graph theory. They also presented an algorithm to minimize the cost based on decision tree.

In 2009, Mahdavi et.al introduced e-CRM in their research and moreover presented an algorithm for fuzzy clustering based on neural networks known as FC-CRM. Their proposed model has been used in e-CRM. They also proposed dynamic clustering through neural network methods as a solution for customer classification. Their model (the same as neural networks) has learning capability.

Finally, it should be noted that numerous theoretical studies have been done in the field of knowledge management and data mining (due to today's knowledge-based approach);however there have been few researches done on costumer attraction prediction, potential customers' identification, and customer purchase basket analysis. This is because knowledge-based issues are not properly institutionalized in organizations and it is rather difficult to make these systems applicable. Furthermore, there is fear of failure of application of these systems.

3. Research Methodology

Identification of variables and sub- variables

In order to present the prediction model, following steps were carried out in an 18-month period March 2008 to September 2009 in the studied organization (furniture producer with 35 years record in this field) based on data mining steps; further details will be presented on each step later:

1. Data Collection,
2. Data Preparation,
3. Data Coding,
4. Determination of target variable,
5. Determination of prediction model variable,
6. Determination of variable coefficients and value of each sub-variable,
7. Numerical use of obtained model,
8. Formation of Decision Tree and extraction of Prediction Optimal Decision Tree
9. Model Evaluation.

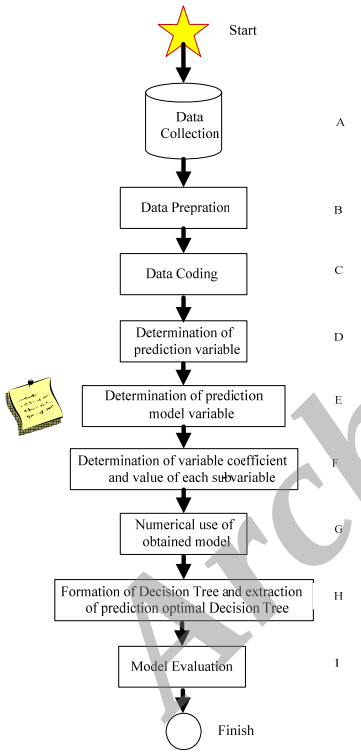

Fig.1. Research methodology

3.1. Data Collection

3.1.1. Variables and sub-variables whose data could be collected were identified consulting experts' opinion and available records based on the customers' attendance(phone/visit)in the studied organization(furniture producer with 35 years record in this field):

Gender (male/ female), name of company/ shop/ office/workplace, date of attendance, address, telephone, cell-phone, fax, e-mail, business size (small, medium, large), product introduction (past purchase, catalogue & CD, publications, personnel, exhibition, website, advertising letter, billboard, reference, representative), product request (phone, after presentation of product), sales expert(expert No.1 expert No.2), product type (single- wall partition, two- wall partition, partition attachments, manager desk, specialist desk, office desk, conference table, shelves, chair (imported, domestic production), types of tables, raw materials and other services), design specialist (expert No.1 to expert No. 4), result of the attendance (purchase, dispense), dispense reasons , lead time (LT), installation staff ,invoice payment, discount, transport and installation cost.

3.1.2. Collecting data according to section 3.1.1 by means of required forms, instructions, and regulations.

3.2. Data Preparation

In this stage, obtained data in section 3.1 which were collected in Microsoft Office Access was refined and prepared. (For the parts which were incomplete or incorrect)

3.3. Data Coding

The data was coded compatible with Microsoft Excel after step3.2.

3.4. Determination of Target Variable

In sale process, considering the importance of attendance's result, "Result of Attendance" was selected as target variable.

3.5. Determination of Prediction Model Variables

Two different methods (Clementine [software] and questionnaire) were employed due to following objectives:

3.5.1. Modeling using decision tree, comparison and selection of optimal tree, and determination of model variables: In this stage, data prepared in section 3.5.2which are popular models in building decision tree (By prediction approach) were used in Clementine by four models of "CHIAD", "CRT", "QUEST" and "C5.0" to identify model's variables for the data of March 2008till September 2009.

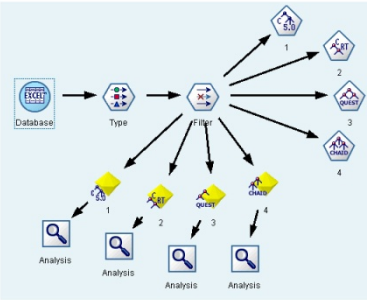

Fig. 2. Modeling for determination of model's variables

3.5.1.1. Data for March 2008 to March 2009

Using data of this year (4240 data in 33920 fields) in the model presented in section 3.5.1, C5.0 provided the most optimal variable and decision tree by 83.96% accuracy which is shown in table1.

Table 1
Importancelevel of variables in C5.0 outputs for the year data

Variable	Importance
Product request	0.411
salesexpert	0.133
Product introduction	0.271
Product type	0.185

3.5.1.2. Data for April 2009 to September 2010

Using data of this period (2080 data in 16640 fields) in the model presented in section 3.5.1, C5.0 provided the most optimal variable and decision tree by 84.62% accuracy which is shown in table 2.

Table 2
Importance level of variables in C5.0 output for data of the first half of the year

Variable	Importance
Product request	0.196
Sales expert	0.132
Product introduction	0.422
Product type	0.25

3.5.2. Using questionnaire to collect experts' opinions

3.5.2.1. Research questionnaire no.1

This questionnaire (Index A) was given to people to identify basic variables in turning potentials into real customers. The output of this survey with the Cronbakh'salpha of74% resulted in identification of the following variables:

i. Product type,
ii. Product introduction,
iii. Sales expert
iv. Product request

3.5.2.2. Research Questionnaire no.2

This questionnaire (Index 2) was given out to people in order to determine the importance of each four variable provided in section3.5.2.1.Using GAHP (After doing independent test in SAS by using Appendix B's output) and the formula $\sqrt[n]{a_{ij1}a_{ij2}...a_{ijn}}$ (1), pair comparisons were summarized in table 3. At this stage of analysis, the Cronbach's alpha was 84% (a_{ijn} opinion of n^{th} person in comparison of variables i^{th} and j^{th})

$$n=1,...,14 \quad i=1,...,4 \rightarrow$$
$$j=1,...,4 \downarrow$$
$$if\ i=j \quad \Rightarrow a_{ijn}=1 \quad (2)$$

Table 3
Importance level of variables in questionnaire output

Variable	Importance
Product request	0.08
Sales expert	0.45
Product introduction	0.28
Product type	0.19

3.6. Determination of variable's and sub- variable's coefficients

3.6.1. Model description

Prediction'starget function and its variables could be defined as follows:

$$Z = A(MR_i) + B(CS_i) + C(MF_i) + D(\sum_{j=1}^{13} KP_j)_i \quad (3)$$

MR_i :Product request of i^{th} customer,

CS_i :Sales expert of i^{th}customer,

MF_i : Product introduction of i^{th}customer and

$(\sum_{j=1}^{13} KP_j)_i$:Product typeofithcustomer.

3.6.2. Determination of coefficients A,B,C&D

In order to determine these coefficients, we used the outputs of sections3.5.1.1, 3.5.1.2, and 3.5.2.2.Table 4 summarizes these outputs.

Table 4
Summary of variable importance level in each output

Variable	Table1	Table2	Table3
Product request	0.411	0.196	0.08
Sales expert	0.133	0.132	0.45
Product introduction	0.271	0.422	0.28
Product type	0.185	0.25	0.19
Degree of Validation	83.96	84.62	84

Following method was used to obtain the value of each

coefficient:

$$v(x_j)=\frac{\sum_{i=1}^{3} w_i x_{ij}}{\sum_{i=1}^{3} w_i} \quad (4)$$

W_t: Correction level of i[th] output, i=1,…,3

X_{ij}: Importance of j[th] variable in i[th]output and j=1,…,4

v (X_j): Value of j[th]variable.

By replacing the valuesof table 4 intoequation (4), equation (5) is obtained:

$$Z=0.23(MR_i)+0.24(CS_i)+0.32(MF_i)+0.21(\sum_{j=1}^{13} KP_j)_i \quad (5)$$

3.6.3. Determination of the value of each sub-variable

Having determined the value of the coefficients of four basic variables, these variables each have sub-variables according to section 3.1.1 To calculate the value of each sub-variable following steps were pursued:

1. Determination of the number of jobs performed byi[th]method (in the proposed period),
2. Calculation of the ratio of the number of jobsperformed byi[th] method to all jobs (in the proposed period),
3. Determination of the number of jobsbyi[th]method which resulted in customer purchase (in the proposed period),
4. Calculation of the ratio of the number of jobsbyi[th]method which resulted in customer purchase to all jobs performed byi[th] method. (In the proposed period),
5. Calculation of mathematical expectation by multiplying sections 2 &4
6. Calculate the mean of the two numbers obtained separately in section 5 for March 2008 to March 2009 and March 2009 to September 2009to get the value of each sub-variable:

$$v(F)=\frac{(A.a)_{2008-2009}+(A.a)_{2009-2010}}{2} \quad (6)$$

Repeating this 6-step method for each sub-variable, we get to table (5).

Table 5
Value of sub- variables using 6- step method

Value of sub-variables	Code	sub- variables	Variables	Value of sub-variables	Code	sub- variables	variables
0.09	P_1	Single-wall Partition		0.3	F_1	Past purchase	
0.06	P_2	Two-wall Partition		0.2	F_2	Catalogue & CD	
0.12	P_3	Partition Attachments		0	F_3	Publication	
0.1	P_4	Management Desk		0.01	F_4	Personnel	
0.11	P_5	Specialist Desk		0.03	F_5	Exhibition	MF_i
0.07	P_6	Office Desk		0.02	F_6	Website	
0.16	P_7	Conference Table	KP_j	0	F_7	Advertising Letter	
0.11	P_8	Shelves		0.12	F_8	Billboard	
0.09	P_9	Imported Chair		0.07	F_9	Reference	
0.08	P_{10}	Domestic Produced Chair		0.06	F_{10}	Representative	
0.09	P_{11}	Types of Tables		0.2	R_1	Telephone	MR_i
0.23	P_{12}	Raw Material		0.42	R_2	Visit	
0.4	P_{13}	Other Services		0.3	F	Expert1	CS_i
-				0.32	M	Expert2	

3.7. Numerical use of obtained model

Using the outputs of stages 3.6.2 &3.6.3 for the data of March 2008 to September 2009,target function (equation 5) is calculated and database for the data of 18 months is created in "Excel" (6320data in 170640 fields).

3.8. Formation of decision tree and extraction of prediction optimal tree of customers

In order to extract prediction optimal tree the following two approaches can be employed:

1. Modeling based on the value calculated for each of

four $0.23(MR_i)$, $0.24(CS_i)$, $0.32(MF_i)$ and $0.21(\sum_{j=1}^{13} KP_j)_i$ variables for each attendance,

2. Modeling based on Z value calculated for each attendance.

3.8.1. Modeling based on value calculated for each of four variables for each attendance and the result of each attendance

By replacing the data of section 3.7 in the model of section3.5.1 based on the assumption of 3.8.1, CHAID provides the most optimal case by 92.25%. (Figs 3,4, table.6)

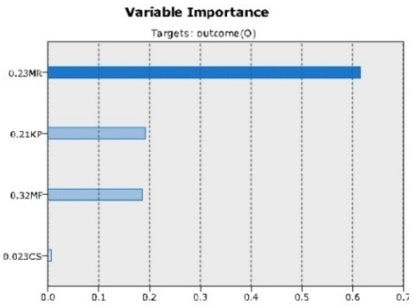

Fig. 3. Importance of variables in the output of CHAID based on four variables

Table 6
Importance of variables in the output of CHAID based on four variables

Variable	Importance
Product request	0.615
Sales expert	0.007
Product introduction	0.186
Product type	0.192

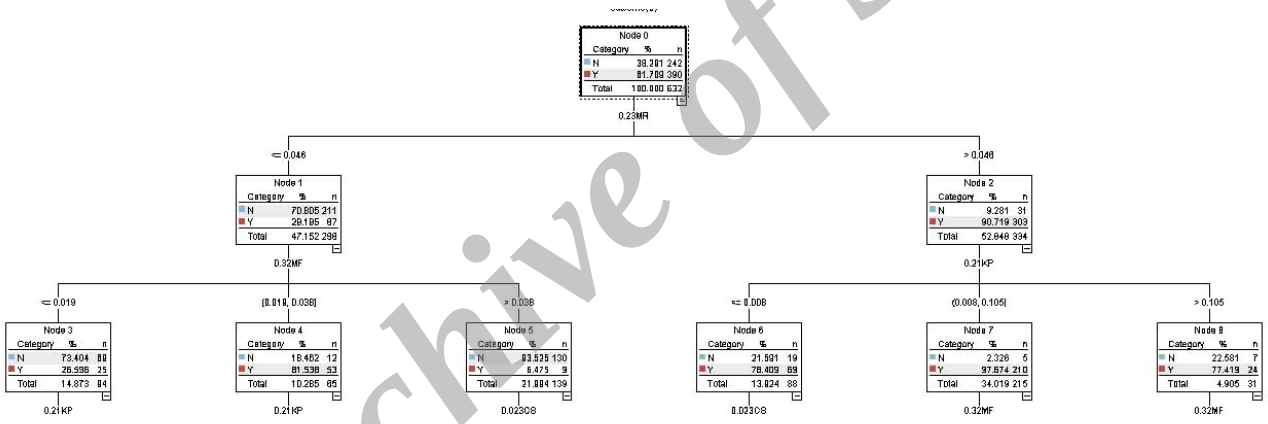

Fig.4. A part of decision tree of important variables of CHAID based on 4 variables.

3.8.2. Modeling based on target function calculated in each attendance and the result of each attendance

By replacing the data of section 7 in the model of section 3.5.1based on the assumption of section 3.8.2, C5.0 provides the most optimal case by 87.03%. (Fig.5&6)

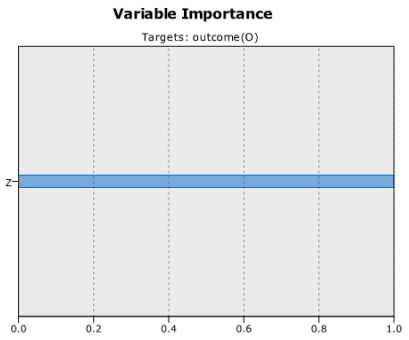

Fig.5. Variables' importance in C5.0 output based on four variables

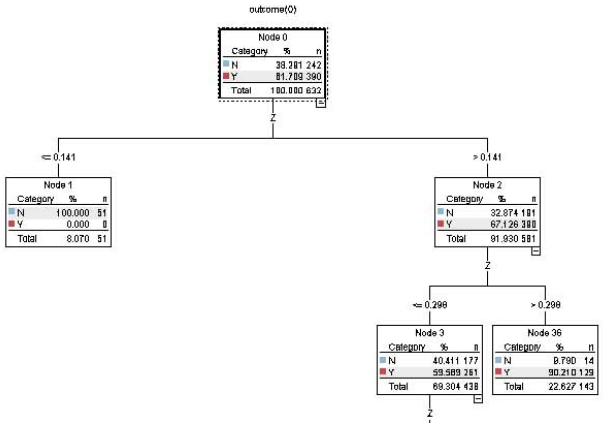

Fig. 6. Part of the decision tree of the important variables of C5.0 based on Z variable.

3.9. Model Evaluation

Information gained from sections 3.8.1 & 3.8.2 was given to the personnel and they were told to use the model in a 5-day period 12th to 17th October, 2009. Results indicate that 16 out of 17 attendances were correctly predicted which is equal to 94%. (11 sells and 5 dispenses, one dispense was because of the absence of sales expert No.1 in the organization and another was because the sales expert couldn't have convinced customer for a personal visit). In time interval of execution and after- performance evaluation it became clear that administrative costs decreased by 8% and transportation costs increased by 5%, while profit of the company increased by 15%[compared to the same number of attendance in the month of July and August, 2009].

4. Conclusion

Combinations of CRM and data mining results in effective responsiveness to customer needs, optimization of capital return and productivity of human force, quality improvement of products, and finally, rapid responsiveness to environmental changes. Therefore, companies seeking to survive in today's highly competitive market and obtain competitive advantage and increase their share of market should change their view from product-orientation to customer-orientation, collect customer data and create customer data store, train staff to use software and hardware systems and effective use of data, provide required technical facilities, identify the requirements of different groups of customers, improve the quality of their products and services.

In the investigated organization decision tree seemed to be a proper tool for identification and classification of the factors for turning potentials into real customers. The four variables Product introduction, product type, sales expert, and the product request are most effective in turning potentials into real customers; The more criteria used in creating decision tree, the more easily customers are identified. The tree obtained based on C5.0 algorithm is closer to field results of section 3.5.1 and performs better in action. In launching final prediction model, personnel tend to more use output information of section 3.8.1.

Another application of decision tree is extraction of rule (Sales staff when telephoning customers, try to persuade them into visiting Show Room of the organization, Sales expert no.1 should be dedicated to the attendants with past purchases and sales expert no.2 to new attendants, attendants who use methods F1,F4,F10 & F8 for attendance must be invested more on, Relations agent should follow decision tree to dedicate attendants to sales expert, considering the dependency between certain products, sales experts must take dependency rules into account).

Appendix A

Research Questionnaire No.1.

Since identifying potential customers can help suppliers and vendors invest on them, attract them through different actions such as advertisement and marketing and turn them into real customers, it's essential to identify these potential customers. Therefore, the purpose of this research is to identify different criteria which enable us to identify potential customers. This questionnaire helps us determine the importance level of several primary criteria and respondents are asked to write down any other criteria they deem to be important. You are required to select one of the options "not important", "less important", "important" and "very important" considering the importance of each criterion.

Appendix B

Research Questionnaire No.2

Identification of potential and real customers is significantly important for any organization. Our organization seeks to better know its customers. In order to predict customer and whether a potential customer turns will into a real customer or not, several criteria are needed. Using field studies in phase-1 and specifying more important criteria we concluded that: size, Product request, sales expert, Product introduction and type of requested product are the most important ones. The purpose of this questionnaire is to compare importance level of these criteria to obtain final weight of each criterion in final model. You are required to determine the importance of each criterion first, using Likert scale (1 to9) in Table1 and then using pair comparison, compare each two criteria in Table2.

Respondent'sinformation:		
First name:	Last Name:	Education:
Occupation:	Date of receiving questionnaire:	

Table A.1

Prediction Criteria					
No.	Criterion	Very important	Important	less important	Not important
1	Business size				
2	Product request				
3	Tendency to get subscription				
4	Having purchase plan				
5	Result of attendance				
6	Market prosperity				
7	Product brand				
8	Goal of customer				
9	Past purchase from other suppliers				
10	lead time				
11	Product introduction				
12	Loyalty level				
13	Capital amount				
14	Business type				
15	Customer appearance				
16	Month of purchase				
17	Past purchase				
18	Product type				
19	Sales expert				
20	Design specialist				

Respondent's information:		
First name:	Last Name:	Education:
Occupation:	Date of receiving questionnaire:	

C1: Product request1 C2: Sales Expert2 C3: Product introduction3 C4: Product type4
Not important:1A little important:3 Relatively important :5 Important:7 Very important:9

Table B.1

9	8	7	6	5	4	3	2	1	C_1
9	8	7	6	5	4	3	2	1	C_2
9	8	7	6	5	4	3	2	1	C_3
9	8	7	6	5	4	3	2	1	C_4

1 .product request could be either personal visit or by phone.
2. Sale experts are 2 people.
3 . product introduction way could be: previous purchase, reference, international exhibition, catalogue &CD, advertisement letters, website, billboard, representative and media.
4. Type of requested products: partition, file, management desk, etc.

Table B.2

Prediction criteria	Product request	Sales Expert	Product introduction	Product type
Product request	1			
Sales Expert		1		
Product introduction			1	
Product type				1

5. References

[1] Baesens, B. Verstreeten, G. Poel, D. (2004). Bayesian network classifiers for identifying the slope of the customer lifecycle of long-life customers. European Journal of Operation Research, 156, 508-523.

[2] Cheng, C.H. Chen, Y.S. (2009). Classifying the segmentation of customer value via RFM model and RS theory. Expert System with Applications, 36, 3, 4176-4184.

[3] Hwang, H., Jung, T. Suh, E., (2004). An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry. Expert Systems with Applications, 26,181-188.

[4] Korach, E., Stern, M., (2008). The complete optimal stars-clustering-tree problem. Discrete Applied Mathematics, 156, 444-450.

[5] Lee, J., H., Park, S., C., (2005). Intelligent profitable customers' segmentation system based on business intelligence tools. Expert system for applications, 29, 145-152.

[6] Mahdavi, B. Shirazi, N. Cho. (2008). Designing evolving user profile in e-CRM with dynamic clustering of Web documents. Data & Knowledge Engineering, 65, 355-372.

[7] Ngai, E. Xiu, L. Chau, D.C.K. (2009).Application of data mining techniques in customer relation management: A literature review and classification. Expert systems with Applications, 36, 2, 2592-2602.

[8] Rygielski, C. Wang, J. Yen, C. (2002). Data mining techniques for customer relation management. Technology in Society, 24, 483-502.

[9] Stefano, C., Sarmaniotis, C. (2004). CRM and customer-centric Knowledge management: an empirical research. Business Process Management Journal, 9, 5, 617-634.

[10] Yong, S., Yoonseong, K. (2008). Searching customer patterns of mobile service using clustering and quantitative association rule. Expert systems with Applications, 34, 2, 1070-1077.