Int. J. Manag. Bus. Res., 2 (1), 41- 50, Winter 2012
© IAU

Customer Retention Based on the Number of Purchase: A Data Mining Approach

[1]S. Mehregan, [2]R. Samizadeh*

[1] *Department of Information Technology, School of Management and Economics, Science and Research Branch, Islamic Azad University (IAU),Tehran, Iran*

[2] *Department of Computer Engineering, School of Computer Engineering, Al zahra University, Tehran, Iran*

ABSTRACT

Purpose: this paper aimed at finding the relationship between the numbers of purchase and the customer's income. The data mining tools were applied in the study to find those customers who bought more than one life insurance policy and represented the signs of good payments at the same time.

Design/ methodology/ approach: in the present research the data mining tools were employed based on CRISP-DM methodology. The K-means algorithm was used for classification and the prediction was based on a proposed formula in Excel worksheet.

Findings: the researcher extracted some simple rules to predict customers' clusters through selecting the customers who bought more than one policy and filtering the income- bringer customers as the companies would be able to use this prediction to change their strategies in relation to different customers.

Originality/value: Utilizing data mining tools to classify different customers in life insurance and prediction based on the classification were new approaches of the study. There was not enough research and implementation in relation to the CRM and data mining in the insurance industry in Iran. Especially CRISP-DM methodology was not used extensively enough in a life insurance investigation.

Keywords: *Cross selling, Data mining, Prediction, Customer retention, CRISP-DM*

INTRODUCTION

Data mining can be defined as the process of selecting, exploring and modeling large amount of data to reveal previously unknown patterns. In the insurance industry, data mining can help firms gain business advantage (Gayle, 2001). It is used in different fields like fraud detection, insurance claim patterns, premium pricing, insurance rate making and finding the customer value for the company (Smith et al. 2000; kiansing and Huan 2001; Bloemer et al., 2003; Cho and Ngai, 2003; Roderick et al., 2004 and Kahane et al., 2007). Moreover, data mining has recently been used frequently in CRM, like classification of customers at risk, creating profitable customers and customer retention (Rygielski et al., 2002; Ryals, 2003; Ngai et al., 2008).

Data mining is actually a part of the knowledge discovery process called KDD. It extracts the knowledge that statistics fails to access. The advantages of data mining are saving time, extracting the knowledge that could

***Corresponding Author, Email:** *Sa.mehregan@gmail.com*

not be gained before, and introducing new solutions (Rygielski et al., 2002). Data mining has been used in insurance in many fields like fraud detection, selection of the sales agents, customer acquisition, finding claim patterns, and so on (Smith et al., 2000; Min and Emam 2002; Yeo et al., 2002; Cho and Ngai 2003; Chen and Hu, 2005; kahane et al., 2007). The application of data mining tools in CRM is an emerging trend in the global economy. CRM is used to enhance the analysis of customer value (Chen and Hu, 2005).Analyzing and understanding customer behaviors and characteristics is the foundation of the development of a competitive CRM strategy (Ngai et al., 2008). A view to CRM says that CRM is a way for Knowledge management, data warehousing and data mining to support the organization in its decision making process (Cunningham et al., 2004). The CRM has four dimensions; customer absorption, customer acquisition, customer retention and development of the customer relationships (Rygielski et al., 2002). Some of the researchers just focused on customer retention either for claim patterns (Smith et al., 2000) or for making good relationship to remain competitive (Min and Emam, 2002) while some focused on the other areas like increasing revenue by the increase in market share or customer acquisition (Yeo et al., 2002).

In the present research is mainly focused on the retaining the customers who bring income to the company and the development of the relationships as it is stated that a 5% increase in customer retention leads to a higher current customer value of 35% to 95% in the companies (Ryals and Knox, 2001). It means that with a right target research the manager can help the company to improve in an appropriate way. Ryals and Knox (2001) also presented a table indicating that by a 5% increase in customer retention for life insurance the company can increase the value of the customer by 84%. It is an interesting ratio for each company to gain; however, good retention in this case may cause the better customer acquisition in future and also better relationship with the current and new customers. Knowing the customers' behavior helps the company to prevent customer defection and increases their retention rate (Song et al., 2004).

RESEARCH METHOD

In the present study CRISP_DM methodology was used to prevent anarchy in the research process. It was first introduced by SPSS, NCR and Daimler Chrysler companies in 1996. CRISP_DM as a standard process is applicable in a variety of fields and projects. This methodology has 6 steps which are business understanding, data understanding, data preparation, modeling, evaluation, and reporting (Chapman et al., 2000). Each step has its own process; business understanding is the phase of finding the research problem and questions. The second phase is data understanding; this phase is in relation with finding research data and inputs. Step three or data preparation refers to the cleansing data, that is to make it ready in the related database and matching the items with what should be the input of the next step. In step four the research is implemented and data mining is executed, the next step is to evaluate the model extracted and the last step is to prepare a complete report for the management (Maalouf et al., 2010).

Business Understanding

The aim of this research was to maximize the profit that customers bring to the company. The customers who purchased more policies would be regarded as the ones with the real potential to bring more income to the company. The researcher wanted to find a way to distinguish those customers who pay regularly; they are referred to as "Income Bringers" in the present paper as the regular payment makes the Life Insurance Portfolio positive and it is very significant for the insurance company. To find a solution the company prepared a database in excel worksheet which presented all the registered information for an underwriter. The data consisted of the insured's personal data as well as the policy information such as age, gender, geographical residence, policy amount, duration, premium payment type, whether the purchaser is the insured person himself or not, and the number of purchases. These attributes are shown in the table 1. These data were extracted from the research literature. Then a number of life insurance experts evaluated the attributes by the use of a questionnaire to determine the importance of each item in

customers' premium payment commitment to the company. The questionnaire had the Likert spectrum for 5 digits to show the results; number 1 stood for the least relationship and number 5 represented the most relationship between the item and the commitment of the customer in payment.

By extracting α>0 .7 and the frequency of all the attributes for more than 3 in Likert spectrum the researcher found that all the attributes have enough validity to be used in the project.

The Proposed Conceptual Model

The process used in the present research is the CRISP-DM method (Chapman et al., 2000) that is known as an inter-science methodology for implementing data mining tasks. The steps and their descriptions are presented in the figure 1.

As the crisp-dm model is somehow complicated the process is defined in a simpler model to show what are the inputs, process and the output of this research. The simple process model is presented in figure 2

Table 1: Attributes ranking based on the influence on the insured commitment

Attributes used for insured commitment in payments	Mean value extracted in Questionnaire in SPSS
Policy duration	4.42
Payment type	5.15
Policy amount	4.89
Number of purchase	4.19
Purchaser	4.31
Age of the insured	4.71
Gender	4.15
Geographical area	4.17

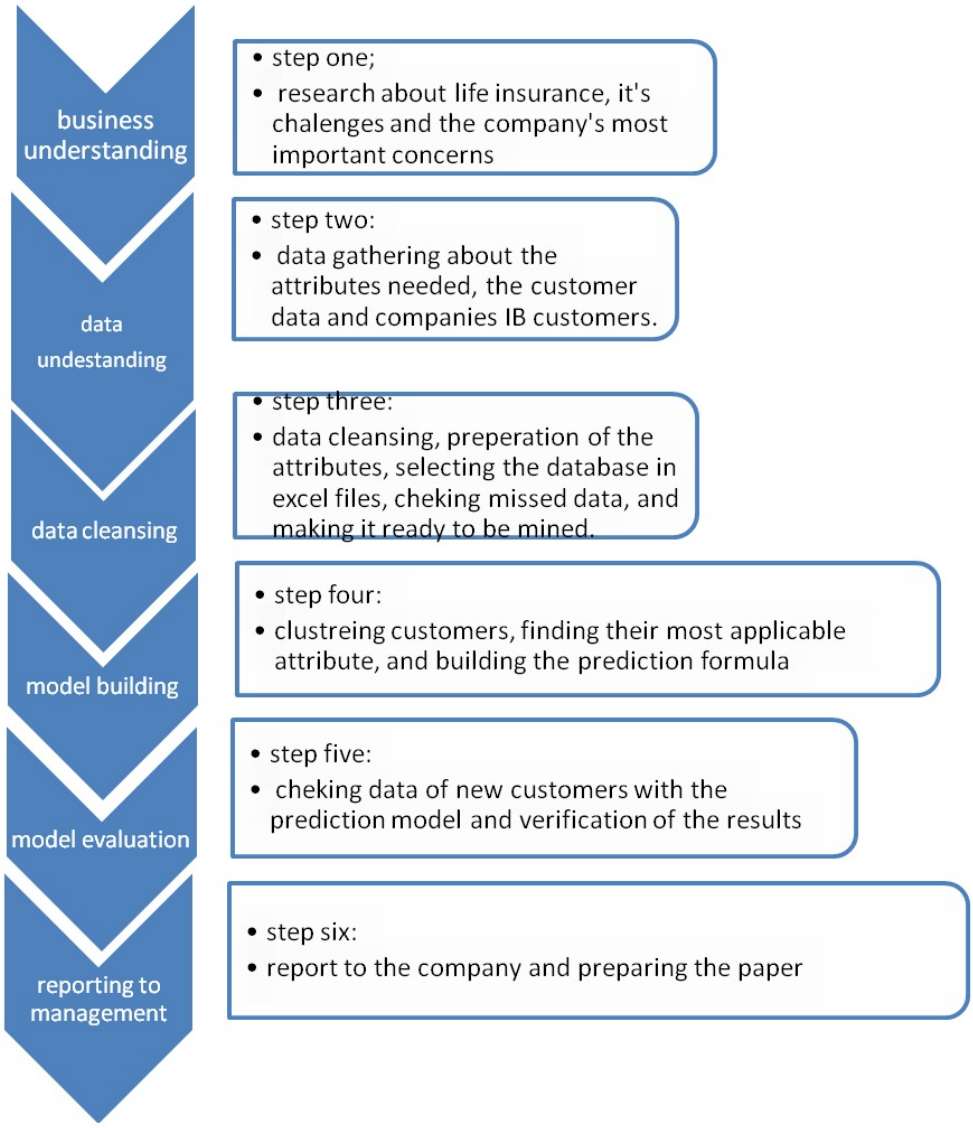

Figure 1: CRISP_DM process for life insurance prediction process

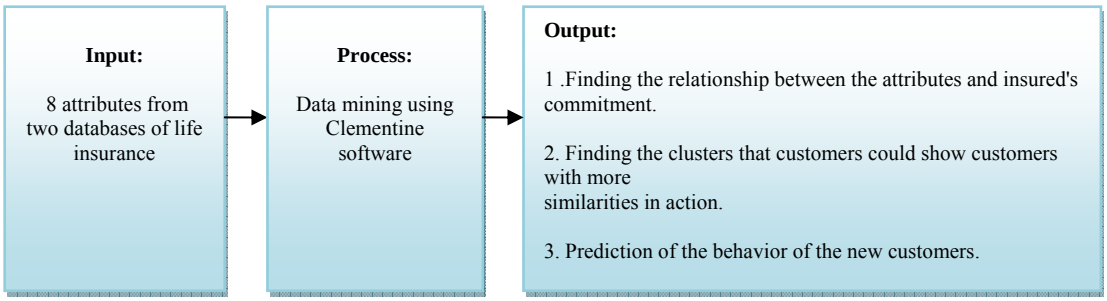

Figure 2: Simple conceptual model

RESULTS AND DISCUSSION

Data was prepared and filtered in Excel worksheet. Data mining process was implemented applying the Clementine version 12.0 software that is data mining SPSS software. First of all an encompassing process was done on the initial data because none of the attributes were ready to be mined in the software. All attributes passed along three to four levels to obtain the prepared data for data mining (step two in CRISP-DM process). The life insurance customers who purchased two or more policies of the company were filtered before importing the data into the Clementine Software.

Segmentation: Data mining started with the segmentation process. K-means algorithm (Maalouf et al., 2010) used to extract three clusters of customers who showed more similarities in their attributes. Then the clusters were studied to find if there would be any rule among the extracted attributes in any cluster. In the final step two levels of prediction on the basis of the information showing the cluster in which the new customer could be classified and the of the customers' commitment were in the hands of company.

The three clusters extracted from the K-means process showed that there would be 4 more important items in the clustering: commitment of the person for premium payment, the gender, the type of premium payment, and the geographical distribution. There was not a noticeable difference in the other attributes: age, policy duration, the capital of the policy and the purchaser. The results of the clustering are presented in the table 2. The best number for clusters was extracted from the auto search of the software. The similarities and the differences among clusters were also studied in this stage; the findings indicated that three clusters released the best data attributes.

The results of analyzing the table by the insurance company experts showed that the majority of the customers not only live in the capital but also have a distribution in the committed areas as compared with the other geographical types; therefore, it was concluded that the people living in the capital were of more value to the company as the customers in the central geographical areas would be mainly distributed in the non commitment clusters.

The same conclusion was true to the annual and monthly payment because the majority of capital customers chose monthly payment, while in the non commitment clusters the customers opted for annual payments in most cases. While clusters one and two showed a stable situation that enabled the company to choose a distinct strategy, cluster three needed further analysis.

Prediction: A table was used to see if the clusters introduced could be helpful in prediction (Smith et al., 2000). The attributes of twenty new customers were studied to see if they match the customer's attributes clusters and it was found out that:

- ✓ An under-18 customer, who lives in the capital and its suburbs, is likely to choose payment annually in the cluster one.

- ✓ An adult, whose policy is purchased by the other person and lives in the central areas, is likely to choose monthly payment in cluster two.

- ✓ A customer, who purchases their own policy and lives in capital or southern regions, is likely to be a man in cluster three.

The formula of such prediction was written in the excel spreadsheet. Each of the three clusters had their own attribute, so finding which customer could be classified in which group was not a difficult process. For example, a male self purchaser would be classified in cluster three.

The extracted rule for a sample formula was as follows:
If: age is under 18, and the city is capital and the payment is monthly;
Then: cluster 1

If: age is above 18, purchaser is other, and the geographical zone is center;
Then: cluster 2

And if: age is above 18, purchaser is self, gender is man, and the geographical zone is capital or south;
Then: cluster 3
The results of the prediction for twenty new customers are given in the table 3.

Table 2: The clusters of the life insurance purchasers based on the customer characteristics

Cluster 1	Cluster 2	Cluster 3
122 people	122 people	109 people
100% children (under 18 purchaser is other person)/ committed 100%/ geographical residence 40% in capital & 25% in southern parts of the country/ premium payment 51% monthly and 22% annually	53% women, 47% men/ committed 0%/ purchaser 100% other person/ geographical residence 33% live in the center of the country/ premium payment more than 75% monthly, less than 1% annually	76% men and other women/ 100% own purchaser/ commitment about 56%/ geographical residence 30% in capital of the country and 22% in southern parts/ premium payment 63% monthly and 12% annually

Table 3: Prediction based on clusters

Number of purhcase	Capital	Inusred gender	purchaser	Age	Policy duration	Geograpical area	Cluster
3	37500000	M	Self	30	20	B	3
2	50000000	F	Self	35	20	D	3
2	50000000	M	Other	1	20	D	1
3	37500000	F	Other	30	20	B	2
3	37500000	M	Other	7	20	B	1
3	50000000	M	Self	33	20	F	3
3	50000000	F	Other	9	20	F	1
2	50000000	F	Other	15	10	C	1
2	50000000	F	Other	6	10	C	1
3	50000000	F	Other	30	20	F	2
3	50000000	F	Other	3	20	F	1
3	85000000	F	Other	24	20	F	2
3	75000000	M	Self	28	20	F	3
2	37500000	M	Other	7	22	B	1
2	37500000	M	Self	36	22	B	3
2	37500000	M	Self	37	18	B	3
2	70000000	M	Self	34	20	B	3
2	40000000	F	Other	37	20	B	2
3	37500000	M	Self	48	10	B	3

The company revealed the results of the premium payments of the customers within nine months. Based on the clusters the researcher expected those in group one to be the committed persons; namely, that customers numbers 4, 6, 8, 9, 10, 12, 15-especially numbers 6 and 15 who lived in the capital city and number 12 who chose the 6-month premium payment- were predicted to be committed in the premium payment, but it proved to be not as expected in regard to customers numbers 6, 12, and 15. Those in the cluster two were expected not to be committed to the on time premium payment. Four customers (numbers 5, 11, 13 and 15) were in cluster three. Two customers (numbers 11 and 13) were from the central cities and numbers 5 and 19 had chosen the monthly payment. There was no focus on any one as two customers showed the signs of nonpayment.

Cluster three needed more attention. The numbers in this group were 2, 3, 7, 14, 16, 17, 18 and 20. The customer number 14 chose the 6-month payment; the customers except number 7 were living in the capital city and expected to be committed to their obligations. The real results showed that three of these people were not committed.

The validation of clustering- based prediction method was 65%. The findings of the study showed that the validation percentage would

increase in case of registering customers' incomes and careers.

Research Appraisal and Model Evaluation

Association rules: based on what was found in the segmentation process and the extracted rule for prediction, the association rules were executed to confirm the findings of the present study. There are two validation processes available to this end: one is the minimum support of the results that means how much the rule depends on the existing data, and minimum support which shows the future probability of the rule (maalouf et al., 2010); minimum support of the rule was set on 10% and minimum confidence was set on the 60% for this research. The results are given in the table 4. Based on these rules it can be said that 85% people who choose the annual payment are the IB ones, and 84% of those who live in the central provinces of the country and choose monthly payment are likely to be non IB purchasers. Moreover, 73% of those who live in the capital and choose quarterly payments are the IB customers. The findings, thus, validated the prediction rules.

CRM Strategies Implementation

The research aimed at designing a method for customer retention. The three clusters out of the segmentation represented different characters. Based on the rules extracted for prediction if the customer shows the characteristics of group1 and wants to purchase more policies; the company can trust him easily and sell more policies. If a customer is in group two, the company should be careful. In this case it is better not to sell more policies to the group two customers, it is better to remind them of the undertaking to manage the previous purchases before applying for another policy. The customers in group three should be checked individually, those with specific signs like annually payment or capital inhabitants does not need that much care, but the company should check the person's file carefully to make decision about selling more policies if the person does not have specific characters which shows his commitment.

CONCLUSION AND FUTURE WORK

Table 2 extracted three clusters of the customers who buy more than one policy from the company. The aim was to find what cluster is the most valuable one in terms of commitment of the customer to premium payment. Cluster 1 and 2 showed explicit situations in which the company can follow straight forward strategies. The company should pay more attention to customers of the capital city and its suburbs and also to those purchasers who want to choose the annual payment for the premium. The target customers of the company can be children, because it seems that parents show more responsibility to pay the children premium than their own or the other policies. But cluster three needs more attention and analysis. Those who located in the cluster three are the self purchaser of their policies. A combination of all the payments and geographical areas are inside this cluster. So it is recommended that the company investigate more on the income level and the career of the customer to find what the most influential items on the customer behavior. Finally the company can use the predictor to implement its strategies on the new customers.

Table 4: Association rules for finding the relationships in life insurance customer's data

Consequent	Antecedent	Support%	Confidence%
IB = yes	Repay type= 12	11.898	85.714
IB = no	Geography= f Repay type= 1	14.731	84.615
IB = yes	Geography= b No. of purchase= 3	17.28	73.77
IB= yes	Geography= b Insured gender= f	15.014	73.585
IB = no	Geography= f Insured gender= f	10.198	72.222
IB = yes	Geography= b No. of purchase= 3 Purchaser= other	12.181	72.993
IB = no	Geography= f	21.153	69.737
IB = no	Geography= f Insured gender= m	11.331	67.5
IB = yes	Geography= b Insured gender= f Repay type= 1	11.331	67.5

REFERENCES

Bloemer, M., Brijs, T., Vanhoof, K. and Swinnen, G. (2003). Comparing Complete and Partial Classification for Identifying Customers at Risk. *International Journal of Research in Marketing*, 20 (2), pp. 117-131.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T. and Shearer, C. (2000). *CRISP-DM 1.0: Step-by-Step Data Mining Guide*. Chapter (2) USA: CRISPDM consortium, 1, pp.13-33.

Chen, Y. and Hu, L. (2005). Study on Data Mining Application in CRM System based on Insurance Trade**.** Proceedings of the 7th International Conference on Electronic Commerce, Xi'an, China, the ACM digital library.

Cho, V. and Ngai, E. W. T. (2003). Data Mining for Selection of Insurance Sales Agents, Expert Systems, 20 (3), pp. 123-132.

Cunningham, C., Song, Y. and Chen, P. P. (2004). Data Warehouse Design to Support Customer Relationship Management Analyses, Proceedings of the 7th ACM international workshop on Data warehousing and OLAP, Washington DC, USA, the ACM digital library.

Gayle, S. (2009). The Business Case for Data Mining in the Insurance Industry: Using Enterprise Mine to Model Pure Premium and Establish Policy Rating Structures, chapter (3), 2 nd ed. USA: SAS Institute Inc., Cary, NC press, pp. 3-12.

Kahane, Y., Leviny, N., Meiriz, R. and Zahavi, J. (2007). Applying Data Mining Technology for Insurance Rate Making: An Example of Automobile Insurance. *Asia Pacific Journal of Risk and Insurance*, 2 (1), pp.33-51.

Kiansing, N.G. and Huan, L. (2001). Customer Retention via Data Mining. *Artificial Intelligence Review*, 14 (6), pp. 569-590.

Maalouf, L. and Mansour, N. (2008). Mining Airline Data for CRM Strategies. *Communications of the ACS*, 1(1), pp. 3-17.

Min, H. and Emam, A. (2002). A Data Mining Approach to Developing the Profiles of Hotel Customers. *International Journal of Contemporary Hospitality Management*,14 (6), pp. 274-285.

Ngai, X-L. and Chau, D. C. K. (2009). Application of Data Mining Techniques in Customer Relationship Management: A Literature Review and Classification. *Journal of Expert Systems with applications*, 36 (1), pp. 2593-2603.

Roderick, M., Rejesus, B. B., Little, A. and Lovell, C. (2004). Using Data Mining to Detect Crop Insurance Fraud: Is There a Role for Social Scientists? *Journal of Financial Crime*, 12 (1), pp. 24-32.

Ryals, L. (2003). Creating Profitable Customers through the Magic of Data Mining. *Journal of Targeting, Measurement and Analysis for Marketing*, 11 (4), pp. 343-349.

Rygielski, Ch., Wang, J-Ch. and Yen, D. C. (2002). Data Mining Techniques for Customer Relationship Management, *Journal of Technology in Society*, 24 (1), pp. 483-501.

Smith, K. A., Willis, R. J. and Brooks, M. (2000). An Analysis of Customer Retention and Insurance Claim Patterns Using Data Mining: A Case Study. *Journal of the Operational Research Society*, 51 (5), pp. 432-541.

Song, H. S., Kim, J. K., Cho, Y. B. and Kim, S. H. (2004). A Personalized Defection Detection and Prevention Procedure based on the Self-organizing Map and Association Rule Mining: Applied to Online Game Site. *Artificial Intelligence Review*, 21 (2), pp.161-184.

Yeo, A. C., Smith, K. A., Willis, R. J. and Brooks, M. (2002). A Mathematical Programming Approach to Optimize Insurance Premium Pricing within a Data Mining Framework. *Journal of the Operational Research Society*, 53 (11), pp.1197-1203.