Int. J. Manag. Bus. Res., 5 (4), 261-268, Autumn 2015

Retaining Customers Using Clustering and Association Rules in Insurance Industry: A Case Study

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

[1] *Department of Industrial Engineering, School of Engineering, Alzahra University, Tehran, Iran*

[2] *Department of Information Technology Management, School of Management and Accounting, Allameh Tabatabaie University, Tehran, Iran*

Received 18 February 2014, Accepted 15 December 2014

ABSTRACT:
This study clusters customers and finds the characteristics of different groups in a life insurance company in order to find a way for prediction of customer behavior based on payment. The approach is to use clustering and association rules based on CRISP-DM methodology in data mining. The researcher could classify customers of each policy in three different clusters, using association rules. At the end of study the characteristics are defined and given to the company, so they could implement CRM strategies based on the newly found differences. Attention to the income and cash earning comes before paying attention to other problems. In most of the companies in developing countries, infrastructural problems of the company like earning enough income prevent the company from effective research implementation on advanced strategies. So this study focuses on basic problems. Utilizing data mining approach to classify customers in life insurance is a new approach among insurance companies in Iran. There are some research in relation to the CRM and data mining, but the contribution of this study is to investigate two new attributes plus those common attributes used before in studying customer behavior, the two attributes are "payment type" and the "purchaser". In order to have a framework, all the process is embedded in CRISP-DM methodology.

Keywords: *Data mining, Life insurance, Customer retention, Decision tree, Segmentation*

INTRODUCTION

a. Understanding the business: Data have always been very important in insurance industry. Nowadays different ways of analyzing data have been introduced to help companies make better decisions in relation to the customer data. Computer and technology made it easier to save, retrieve, and use data. One of different ways that companies use in their day to day analysis is data mining. The use of data mining in insurance industry has accelerated recently in developing countries. Data mining tools can form a model. They take data and construct a representation of reality in the form of that model to help in decision making. The resulting model describes patterns and relationships in the data (Rygielsky et al., 2002). These models are used in different fields like fraud detection (Rejesus et al., 2004), insurance claim patterns,

***Corresponding Author, Email:** *Mehregan921@atu.ac.ir*

premium pricing, insurance rate making and finding the customer value for the company in the insurance field (Smith et al., 2000; Kiansing and Huan, 2001, Bloemer, 2003; Kahane, 2007). In other side, data mining has frequently used in CRM, like classification of customers at risk, creating profitable customers and customer retention (Min and Emam, 2002; Rygielsky et al., 2002; Ryals, 2003; Song et al., 2004; Ngai et al., 2009). CRM consists of four dimensions: 1. Customer Identification; 2. Attraction; 3. Retention; and 4. Development (Ngai et al., 2009). Segmentation, and ranking of customers based on tendency to buy, order frequency, and purchasing behavior is called the customer acquisition process. Segmentation is to separate customers into groups according to common characteristics so that marketing and operational strategies can be targeted to specific populations (Maalouf and Mansour, 2008). For the customer identification and retention, association rules and classification proved to be appropriate methods (Ngai et al., 2009). The study continues by literature review, and then the CRISP-DM methodology is introduced as the base method of the study. Then data is analyzed and explained in results and conclusion part.

Literature Review

There is literature about CRM strategies based on data mining and some models like SAS to determine the level of data mining match with insurance companies (Chen and Hu, 2005). Some investigations are also tried to extract health insurance information systems (Ngai et al., 2009). Some has worked on premium pricing and a pattern for call centers and also for claim patterns (Smith et al., 2000; Yeo et al., 2002). Some researched on selection of sales agents (Cho and Ngai 2003). Loots and Grobler found CRM the best way to retain customers, they combined CRM with public relations and made a model to research on short-term insurance customers and brokers, they found three inconsistencies between what companies think necessary for customer retention and what customers want. Customer mostly cares the quality of relationship, the price and the benefit of product but the companies think that customer is more influenced by the price, service levels and then the quality of relationship. This inconsistency between perceptions makes the retention unsuccessful (Loots and Grobler, 2014). They also focused on IT as a good retention enabler in insurance companies (Loots and Grobler, 2014). IT-enabled CRM has changed the shape of the relationship between firms and Clients, so companies can discover and manage customer Knowledge (Karakostas et al., 2004). From the technological view to CRM there are three phases in CRM lifecycle: 1.the integration of customer data, 2.the analysis for deeper understanding and 3.action to provide a positive impact (Karakostas et al., 2004). Karakastas et al. (2004), found that the companies do not use CRM to their full capacity and they don't implement CRM as a strategy, although 95% of them told they regard CRM as competitive advantage (Karakostas et al., 2004). They also found there were no technical obstacle to implement CRM but there is a lack of change management to do it (Karakostas et al., 2004). Shim et al. (2004), researched on CRM strategies to find VIP customers and their purchase patterns in an e-business. To do so they used data mining techniques such as regression and decision tree. (Shim et al., 2012). They found that some customers are responsible for 70% of the company's revenue. Using decision tree C4.5 they found these customers are those who buy some different items at once (Shim et al., 2012). Bull concluded that even the insurance company under his study was using an expensive CRM system, but the customers were unsatisfied. They amended the system and used a better data processing to gain useful knowledge and trained the intermediaries so they could improve customer satisfaction (bull, 2010). Other studies also worked on different subjects focused on insurance companies like fraud detection (Rejesus et al., 2004), insurance claim patterns (Smith, et al., 2000) premium pricing (Yeo et al., 2000), and insurance rate making (Kahane et al. 2007).

As part of the data mining investigations in insurance companies, this study wants to classify the customers in life insurance department to execute some strategy for customer retention and finally increase the profitability of the organization using data mining analysis. The profitability here means "the cash flow customers bring to the company by regular premium payment". Two

kinds of life insurance packages called “life and capital insurance” (L&C) and “life and savings insurance” (L&S) are used in this case study. Although they are different, but some common characteristics exist between them that made the comparison possible. So this study analyses the customers in two different life packages. Then compares them to find common characteristics of different customers of the company and find a way to make them more profitable. This study is somehow a continuation of a study named "Customer Retention Based on the Number of Purchase". Data and research's base are the same but this study's focus is on classification of all customers. The previous study mostly focused on those customers who purchased more than once from the company to extract a pattern of their behavior (Mehregan and Samizadeh, 2012).

RESEARCH METHOD

“First step in the research is to match the problem with data mining solution, because not every problem could be solved with the data mining. Some directions to use data mining are as follows: the potential of having great impact, large amount of data accessible and few missing data, the relationship between attributes and enough information about the dimensions of the research.”(Kiansing and Huan, 2001) In order to keep in a structured format this study follows the standard of CRISP-DM methodology in data mining (Chapman et al., 2000). Based on this method the study follows six steps, which the first step, the business understanding, was briefly described in the introduction. The second step is finding attributes; which is done by the attribute selection among different fields of data, using the research field’s literature. The third step is to cleanse data and make it ready to be mined. It is an important step in data mining, because bad cleansing may make data unreliable for more investigation. The forth step is to build the model, and evaluate it in the next step and finally make a report to the management.

The Proposed Conceptual Model

The conceptual model of the research is based on the CRISP-DM methodology, presented in figure 1.

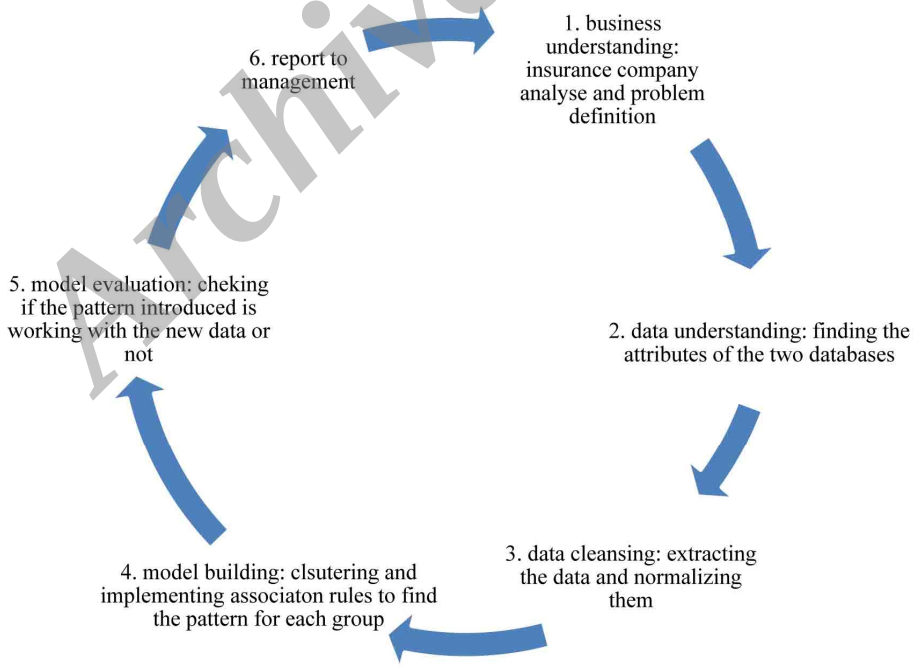

Figure 1: The research process based on the CRIPS-DM method

B and C. Understanding and Cleansing Data: The process in this research is implemented in Clementine software. Some operations like data cleansing was based on the expert's consultation and some other like checking the attribute's validity are done in SPSS software. The database of L&C policy consisted of the 39 fields for 2680 records within three consequent years. The fields of data were the detailed information that company asks any insured. After some time another database was given to the researcher which showed the process of the premium payments to find the level of cash payment of the customers. Within these 39 fields with so much discussion and analysis with experts 8 fields were selected. Among these 8 fields 6 have been used before in the literature (Mehregan and Samizadeh, 2012), but two of them are new, "the number of purchase" and "the purchaser". The number of purchase shows that how many policies have been sold to the members of a family or employees of a company. And the purchaser for a special record shows that did the person buy the policy himself or someone else has bought it.

The L&S policy contained 32 fields for 1268 records which so many of fields were in common with the L&C policy. But only 9 attributes extracted from all of these. Name of the attributes and their mean rank for checking the importance of effect on customer's behavior, in freedman's test in SPSS are given in table 1.

For the L&S finding the geographical area was too difficult because it was not registered in most of the insurer's data. The researcher could not complete it as an attribute, so it would be eliminated from the database of the L&S. The reason for not registering the data in L&S was that most of the L&S customers are the companies who buy the insurance for an employee. So the address was the companies address not the insured person's.

Except the geographical area all the data are common in the two databases and will be used then in the insurance data mining process of the M company.

RESULTS AND DISCUSSION

Association Rules: In data mining, association rule is a popular and well researched method for discovering interesting relations between variables in large databases (Hajizadeh et al., 2010). The most well-known algorithm to discover association rules is the "Apriori Algorithm". It was introduced by Agrawal and Srikant in 1994 (Bloemer et al., 2003). The researcher used the association rules to make sure if there is any kind of relationships between the attributes to support the company in decision making.

For L&C policies, the Apriori algorithm (with 77% support) showed that those who pay premium on a yearly base, are the IB ones. The next rule was that with 73% support those who live in the center area of the country are not IB customers. With 76% support we can say the people in the capital of Iran and the southern parts are income bringers (IB), so this rule supports what we get from the tree. With 74% support those who bought three policies are IB and with 64% those who bought a policy based on six months payment are the IB ones too; so we can say the income increases by the persons who buy policies for a longer time base of payments. Results are presented in table 2.

Table 1: Attributes classification

Attribute type of investigations	Attributes or independent variables		Mean rank
Investigated before in several literature	1.	Policy duration	4.42
	2.	Policy amount	4.89
	3.	Number of purchase	4.19
	4.	Age of the insured	4.71
	5.	Gender	4.15
	6.	Geographical area	4.17
First time to investigate	7.	Payment type	5.15
	8.	Purchaser	4.31

The association rules for the L&S with 82% confidence showed if the purchaser chose a monthly payable payment, the customer is not the IB one. And with 80% confidence "companies" are not income bringers to the company. The associated results are presented in table 3.

Segmentation: The researcher used K-means algorithm to segment the policyholders in three groups. The results are presented in the table 4. It shows the biggest group of buyers with 62% of the share. This group differs from the other groups in two fields; "number of purchase" and "payment type".

Table 2: Association rules for L&C policy

Consequent	Antecedent	Support %	Confidence %
IB=yes	Payment type= yearly	12.134	77.181
IB=no	Geography= F Payment type=1.0	13.762	73.373
IB=no	Geography= F Number of purchase= 1.0	15.309	70.745
IB=yes	Payment type= 6.0	10.342	64.567
IB=no	Sex= female Payment type= 1.0	14.984	64.13
IB=no	Payment type=1.0 Number of purchase=1.0	44.218	60.958
IB=no	Purchaser= other Payment type=1.0 Sex= male	10.261	60.317

Table 3: Association rules for L&S policy

Consequent	Antecedent	Support %	Confidence %
IB=no	Payment type=1.0	84.502	82.625
IB=no	Sex= male Payment type= 1.0	64.763	82.368
IB=no	Purchaser=other	63.458	80.463

Table 4: Three group classification for L&C policy

Groups	Percentage of the total buyers	Age	capital	Policy duration	Number of purchase	Repayment type
1	62%	26	61000	20 years	1	1, 3 monthly
2	23%	22	67000	21 years	2, 3	1, 3 monthly
3	15%	25	65000	21 years	1, 2	Yearly

The results of L&S buyers' segmentation presented in table 5.

The results of the segmentation are compared in the table number 5.

Comparing the three groups of the two policies shows that generally the L&C buyers are about 5 years younger than L&S buyers. The amount they choose is lower, and the policy durations are higher. The two groups are common in the payment types and the number of personal purchases. Comparing results are presented in table 6.

d. model building: Decision trees are powerful and popular tools for classification and prediction (Hajizadeh et al., 2010). Here the prediction is the most important feature for model building in this study. The model used for classifying the customers in two databases was the C5 tree. C5 is the next version of the C4.5 classification tree algorithm by Quinlan (1993). C5 constructs classification trees by recursively splitting the instance space into smaller subgroups until the subgroup contains only instances from the same class (a pure node),or the subgroup contains instances from different classes (impure) but the number of instances in that node is too small to be split further (Bloemer et al., 2003). The researcher used ten-fold cross validation method to prevent over fitting in the model. Over fitting pertains to the phenomenon where one gets a very good fit on the data which is used to build the model, but poor fit when the model results are applied on a new set of observations (Kahane, 2007). The results of the process are represented in two ways. One is the model confidence and the other one is the coincidence matrix.

Table 5: Three group classification for L&S policy

groups	Percentage of the total buyers	Age	capital	Policy duration	Repayment type	Number of purchase
1	74%	30	69000	12 years	1, 3 months	Companies or 1(personal)
2	20%	25	97000	15 years	1, 6 months	1, 2
3	6%	28	80000	12 years	6 month or yearly	1(personal) or some companies

Table 6: Comparison of the three group classifications

groups	Attributes	L&C	L&S
1	Duration Payment type Number of purchase	62% 20 years 1, 3 monthly 1	74% 12 years 1, 3 months Companies or 1(personal)
2	Duration Payment type Number of purchase	23% 21 years 1, 3 monthly 2, 3	20% 15 years 1, 6 months 1, 2
3	Duration Payment type Number of purchase	15% 21 years Yearly 1, 2	6% 12 years 6 month or yearly 1(personal) or some companies

Table 7: Sample of prediction for L&C policy

Number of Purchase	Capital	Sex	purchaser	Age	Policy duration	Geography	Payment type	IB	Partition	C-IB	CC-IB	Real data
1.00	38000	M	Other	1	30.00	B	3.00	Null	1-training	Yes	0.621	Yes
1.00	100000	M	Self	26	20.00	B	12.00	Null	2-tesiting	Yes	0.802	Yes
1.00	120000	F	Self	48	15.00	C	1.00	Null	2-testing	No	0.765	No
1.00	38000	M	Self	24	15.00	E	1.00	Null	1-training	No	0.667	Yes

By choosing the validation test method, the decision tree making will start. The data class is the premium payment by the customers and the others are the attribute ones. the confidence rate for the L&C is calculated as 74% percent and the coincidence matrix showed 88% and 30% for the correctly classified TP (true positive) rate which considered as the highest results that the researcher gained from different models.

e. Model evaluation: After finding the segments and introducing C5 tree to predict the profitability of customers to the company, twenty new data of recent customers was given to the model to predict if they would be income bringers or not. After three months, predicted data was compared with real data. Finally decision three showed enough reliability for predictions. The sample predictions are given in table 7.

Reliability of the model based on 20 new data was about 86% for the L&C policy model and 90% for the L&S policy.

CONCLUSION

F. Report: The findings of association rules in this research showed that two important attributes for calculating profitability of customers are "payment type", the "purchaser" and somehow "geography". It has been shown that most of L&C buyers are profitable for the company and most of the L&S customers (companies) are not profitable. It's the most important difference between the two policies in the company. As a whole people in segments of L&C have chosen longer times for payment, while family purchasers mostly have chosen one purchase, based on what we saw in segmentation process. At the end, the C5 tree could predict the profitable customers to the company with a high level of confidence.

Future Works

A missing factor in this research is the "wealth amount" of the customers under study. This factor could be presented by the customer's monthly "income". Based on the results of the research, researchers found that so many of the behaviors seem to have a relationship with income factor. If the company registers the income of the purchaser and develops the research again, it hopefully can gain better results. So researchers suggest new researchers to study the relationship between the income levels of customers with their profitability for the company and on time payments.

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.

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.

Bull, C. (2010). Customer Relationship Management (CRM) Systems, Intermediation and Disintermediation: The Case of INSG. *International Journal of Information Management*, 30 (1), pp. 94-97.

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) 1,

USA: CRISPDM Consortium, 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.
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), 2nd ed. USA: SAS Institute Inc., Cary, NC press, pp. 3-12.
Hajizadeh E., Davari Ardakani H. and Shahrabi J. (2010). Application of Data Mining Techniques in Stock Markets: A Survey, Industrial Engineering Department, Amirkabir University of Technology, Tehran, Iran.
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.
Karakostas, B., Kardaras, D. and Papathanassiou, E. (2005). The State of CRM Adoption by the Financial Services in the UK: An Empirical Investigation. *Information and Management*, 42 (6), pp. 853-863.
Kiansing, N. G. and Huan, L. (2001). Customer Retention via Data Mining. *Artificial Intelligence Review*, 14 (6), pp. 569-590.
Loots, H. and Grobler, A. F. (2014). Applying Marketing Management and Communication Management Theories to Increase Client Retention in the Short-Term Insurance Industry. *Public Relations Review*, 40 (2), pp. 328-337.
Maalouf, L. and Mansour, N. (2008). Mining Airline Data for CRM Strategies. *Communications of the ACS*, 1 (1), pp. 3-17.
Mehregan, S. and Samizadeh, R. (2012). Customer Retention Based on the Number of Purchase: A Data Mining Approach. *International Journal of Management and Business Research*, 2 (1), pp. 41-50.
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.
Rejesus, R. M., Little, B. B., Lovell, A. 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.
Shim, B., Choi, K. and Suh, Y. (2012). CRM Strategies for a Small-Sized Online Shopping Mall Based on Association Rules and Sequential Patterns. *Expert Systems with Applications*, 39 (9), pp. 7736-7742.
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.