International Journal of Information Technology and Knowledge Management
January-June 2008, Volume 1, No. 1, pp. 51-56

DATA MINING TECHNIQUES FOR IDENTIFYING THE CUSTOMER BEHAVIOR OF INVESTMENT IN LIFE INSURANCE SECTOR IN INDIA

Kanwal Garg, Dharminder Kumar & M. C. Garg

Abstract

Establishing a data warehouse of customer data and analyzing customer behavior have helped companies across different industries to improve their bottom line significantly. In early nineties, banking sector revolutionized the credit card industries by building its entire business around informed intelligence – by collecting customer driven information and by applying predictive modeling technique of Data Mining. But, in contrast there has been little effort made for the Life Insurance Sector in India to repeat the earlier success as achieved in banking sector. Therefore in this paper we have given more emphasis on identifying the trend of customer investment behavior in life insurance sector in India. To achieve the said objective we will use data mining techniques such as clustering – descriptive data mining technique and decision tree analysis – predictive data mining technique.

Keywords: Data Mining (DM), clustering, decision tree.

1. Introduction

Because of the rapid progress of information technology, the amount of information stored in insurance databases is rapidly increasing. These huge databases contain a wealth of data and constitute a potential goldmine of valuable business information. As new and evolving loss exposures emerge in the ever-changing insurance environment, the form and structure of insurance databases change. Finding the valuable information hidden in those databases and identifying appropriate models is a difficult task.

Data mining (DM) is the process of exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns and rules. A typical data mining process includes data acquisition, data integration, data exploration, model building, and model validation. Both expert opinion and data mining techniques play an important role at each step of this information discovery process.

This paper introduces two important data mining techniques for identifying the customer behavior of investment in life insurance sector in India i.e. cluster discovery methods and decision tree analysis.

2. DATA MINING PROCESS

Its primary goal is to extract knowledge from data to support the decision-making process. Two primary functions of data mining are **"**prediction**"** which involves finding unknown values/relationships/patterns from known values and "description", which provides interpretation of a large database. A data mining process generally includes the following four steps.

Step 1: Data acquisition. The first step is to select the types of data to be used. Although a target data set has been created for discovery in some applications, DM can be performed on a subset of variables or data samples in a larger database.

Step 2: Preprocessing data. Once the target data is selected, the data is then preprocessed for cleaning, scrubbing, and transforming to improve the effectiveness of discovery. During this preprocessing step, developers remove the noise or outliers if necessary and decide on strategies for dealing with missing data fields.

Step 3: Data exploration and model building. The third step of DM refers to a series of activities such as deciding on the type of DM operation; selecting the DM technique; choosing the DM algorithm; and mining the data to extract novel patterns hidden in databases.

Step 4: Interpretation and evaluation. The fourth step of the DM process is the interpretation and evaluation of discovered patterns. This task includes filtering the information to be presented by removing redundant or irrelevant patterns, visualizing graphically or logically the useful ones, and translating them into understandable terms by users.

3. CREATION OF CUSTOMER DATABASE

Data un-availability can pose significant problems in conducting appropriate analysis. In order to get true benefit for the study of life insurance sector in India, data needs to be collected from all type of Life insurance customers covering socio-economic factors like: occupation, income group, age group, educational competency, marital status, residential location (rural/urban) etc. as independent variable and 39 dependent variables to cover up the following objectives through close ended questionnaire:

Trend of Investment

We will incorporate such questions which will help us to identify the trend of investor/ customer in life insurance policies as well as in other financial sectors such as mutual funds/shares/debentures, FDs with banks and in other postal savings as regulated in India.

Customer Acquisition

Traditionally, life insurance companies uses the services of only brokers as the channel to acquire customer, but today leadership/ goodwill of the company, initiation/ selling skill of agent, references (friend/relatives), media, product range, sms service on mobile etc. helps us to acquire more customers.

After data has been properly gathered, cleaned and stored, and an analytics process created, it is necessary to implement the data mining techniques.

4. Data Mining Techniques in the Life Insurance Sector of India

Data mining methodology can often improve existing models by finding additional important variables, by identifying interactions, and by detecting nonlinear relationships. DM can help life insurance companies to make crucial business decisions and turn the new found knowledge into actionable results in business practices.

Customer Level Analysis

Using the Associated Discovery DM technique, insurance firms can more accurately select which policies and services to offer to which customers. With this technique insurance companies can Segment the customer database to create customer profiles and perform sequential market basket analyses on customer segments.

Database segmentation and more advanced modeling techniques enable analysts to more accurately choose whom to target for retention campaigns. A logistic regression model is a traditional approach to predict those policyholders who have larger probabilities of buying the life insurance products prevailing in India. Identifying the target group for retention campaigns may be improved by modeling the behavior of policyholders.

Developing New Product Lines

Insurance firms can increase profitability by identifying the most lucrative customer segments and then prioritize marketing campaigns accordingly. Problems with profitability can occur if life insurance companies do not offer the "right" policy or the "right" rate to the "right" customer segment at the "right" time. With DM operations insurance firms can now utilize all of their available information to better develop new products and marketing campaigns.

4.1 Clustering – Descriptive Data Mining Technique

Clustering is one of the most useful tasks in data mining process for discovering groups and identifying interesting distributions and patterns in the underlying data. Clustering problem is about partitioning a given data set into groups (clusters) such that the data points in a cluster are more similar to each other than points in different clusters[12].

Clustering methods perform disjoint cluster analysis on the basis of Euclidean distances computed from one or more quantitative variables. Thus, we can specify the clustering criterion that is used to measure the distance between data observations.

When classifying different kind of samples a way to represent the sample in a mathematical way is needed. From now on we assume that the features are represented in a ***feature vector***. A feature vector is a vector including different features for the sample. That is, with l features xi the feature vector is of the form

$$\boldsymbol{x} = [x1, x2, \ldots, x\text{l}]T$$

Where T denotes transposition and xi are typically real numbers.

Definition of a Cluster

Let us define some basic concepts of clusters in a mathematical way. Let X be a set of data, that is

$$X = \{\boldsymbol{x}1, \boldsymbol{x}2, \ldots, \boldsymbol{x}n\}.$$

A so called *m-clustering* of X is its partition into m parts (clusters) $C1, \ldots, Cm$, so that

1. None of the clusters is empty; $Ci \neq \emptyset$
2. Every sample belongs to a cluster
3. Every sample belongs to a single cluster (crisp clustering); $Ci \cap Cj = \emptyset$, $i \neq j$

Naturally, it is assumed that vectors in cluster Ci are in some way "more similar" to each other than to the vectors in other clusters. Figure 1 illustrates a couple of different kind of clusters; compact, linear and circular.

4.2 Decision Tree – Predictive Data Mining Technique

A decision tree is a predictive model that, as its name implies, can be viewed as a tree. Specifically each branch of the tree is a classification question and the leaves of the tree are partitions of the dataset with their classification. The decision tree is efficient

Figure 1: Couple of Different kind of Clusters

and is thus suitable for large data sets. Decision trees are perhaps the most successful exploratory method for uncovering deviant data structure. Trees recursively partition the input data space in order to identify segments where the records are homogeneous.

For instance if we were going to classify customers who on the basis of their nature job profile , their income level opt for a specific Life insurance policy in India. This decision tree may provide a prediction for a kid of policy the customer generally buy, the maturity value they generally get either on due date or before due date.

This prediction may help the Life insurance service provider to make a target to specific income group customer to buy their newly launched policies. Another way that the decision tree technology has been used is for preprocessing data for other prediction algorithms. Specific decision tree methods include Classification and Regression Trees (CART [4]) and the count or Chi-squared Automatic Interaction Detection (CHAID) algorithm. CART and CHAID are decision tree techniques used to classify a

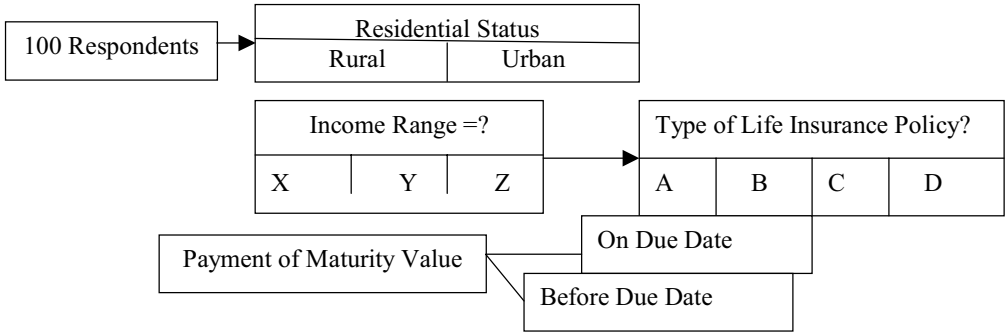

Figure 2: Shows a Decision Tree as a Predictive Model that Makes a Prediction on the Basis of a Series of Decision for 100 Respondents

data set. The CHAID method of tree construction specifies a significance level of a Chi-square test to stop tree growth.

CONCLUSION

The key to gaining a competitive advantage in the life insurance industry in India is found in recognizing that customer databases, if properly managed, analyzed, and exploited, are unique, valuable corporate assets. Insurance firms can unlock the intelligence contained in their customer databases through modern data mining technology. Data mining uses predictive modeling, database segmentation, market basket analysis, and combinations thereof to more quickly answer crucial business questions with greater accuracy. New products can be developed and marketing strategies can be implemented enabling the insurance firm to transform a wealth of information into a wealth of predictability, stability, and profits. The life insurance companies in India may focus on that segment of customers from where the maximum value of insurance policies can be captured. On the basis of these results they may revise their existing business models for targeting the customer segment, setting the prices of their product, designing of life insurance policies etc.

REFERENCES

[1] Adya, M. (1998), "How Effective Are Neural Networks at Forecasting and Prediction? A Review and Evaluation." *Journal of Forecasting*, **17**, 481-495.

[2] Berry M. A. and G. S. Linoff, (2000), Mastering Dam Mining. New York, NY.: Wiley.

[3] Bishop C. M., (1995), Neural Networks for Pattern Recognition, New York: Oxford University Press.

[4] Breiman L., J. H. Friedman, R. A. Olsben, and C. J. Stone, (1984), Classification and Regression Trees. New York, NY.: Chapman & Hall.

[5] Borok, L. S. (1997), "Data Mining: Sophisticated Forms of Managed Care Modeling Through Artificial Intelligence." *Journal of Health Care Finance*, **23**(3), 20-36.

[6] Carpenter G. and S. Grossberg. (1988), "The Art of Adaptive Pattern Recognition by a Self-organizing Neural Network." *IEEE Computer*, **21**(3), 77-88.

[7] Cheesman, P. (1996), "Bayesian Classification (AutoClass): Theory and Results," in Advances in Knowledge Discovery and Data Mining, (ed.), by Fayyad U., G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. CA: AAAI Press/The MIT Press, 153-180.

[8] Ester M., H. Kriegel, J. Sander, and X. Xu. (1998), "Cluttering for Mining in Large Spatial Databases." Special Issue on Data mining, *KI-Journal*, **1,** Scien Tec Publishing.

[9] Fayyad U. M., G. Piatetsky-Shapiro, P. Smyth and R. Uthumsamy (Eds), (1996), *Advancesin Knowledge Discovery and Data Mining*. Cambridge, MA: The MIT Press.

[10] Fisher D., M. Pazzani and P. Langley, (1991), Concept Formation: Knowledge and Experience in Unsupervised Learning. San Mateo, CA: Kanfmann , J.R. 1983, "Induction of Decision Trees." *Machine Learning*, **1**, 81-106.

[11] Goldberg D. E., (1989), Genetic Algorithms in Search, Optimization and Machine Learning.Morgan Kaufmann.

[12] Guha S., R. Rastogi and K. Shim K., (1998), "CURE: An Efficient Clustering Algorithm for Large Databases." Proceedings of the ACM SIGMOD Conference.

[13] McClelland and the PDP Research Group, Cambridge, MA: The MIT Press, 318-362.

[14] SAS Institute. (2000), Enterprise Miner, Cary, N.C.: SAS Institute.

[15] Tufte E. R., (1983), The Visual Display of Quantitative Information, Graphics Press, Cheshire, CN.

Kanwal Garg
Asstt. Prof., MMICT & BM, Mullana (Ambala)
Research Scholar, Deptt. of Computer Science and Engineering
GJUS & T, Hisar (Haryana), INDIA
E-mail: gargkanwal@yahoo.com; gargkanwal@gmail.com

Dr. Dharminder Kumar
Professor, Deptt. of Computer Science and Engineering
Guru Jambheshwar University of Science and Technology
Hisar (Haryana), INDIA

Dr. M. C. Garg
Reader, Haryana School of Business
Guru Jambheshwar University of Science and Technology
Hisar (Haryana), INDIA