Decision Tree induction based classification for mining Life Insurance databases

S.Balaji[1] ,Dr.S.K.Srivatsa[2]

[1] Research Scholoar,Vels University Email: srisaibalaji@rediffmail.com

[2]Senior Professor, St.Joseph Eng.College Chennai-600100

Abstract- Data Mining may be viewed as automated search procedures for discovering credible and actionable insights from large volumes of high dimensional data.This study examine the characteristics of the Decision Tree Induction technique how they can be used to mine the insurance database for predicting the customer preferences over the life insurance policies.The proposed method adopted in this paper for segmentation of customer utilizes decision tree technique for customer preferences towards products.

Keywords-Data mining,Insurance,Decision Tree.

1.INTRODUCTION

Life insurance is an appropriate financial tool for managing and mitigating the financial risk associated with untimely death. However, Life Insurance decisions are often complex. The choice of a life insurance product for an Indian Consumer is now a problem of plenty.Insurance industry in India is aims to protect the interest of and secure fair treatment to policyholders and to bring about speedy and orderly growth of the insurance industry (including annuity and superannuation payments), for the benefit of the common man, and to provide long term funds for accelerating growth of the economy. At the IRDA, the regulator's goal is to see that life insurers are increasingly able to attract, motivate and retain outstanding people, committed to adopting a 'needs-based' approach to financial advice.With Datamining operations insurance firms can now utilize all of their available information to better promote the products and guide the potential customer for policy preference decision.

II.LITERATURE SURVEY

Datamining can help insurance firms make crucial business decisions and turn the new found knowledge into actionable results in business practices such as product development, marketing, claim distribution analysis, asset liability management and solvency analysis. Data mining technology can filtrate and classify customer resources of insurance, divide credit customers into several grades, to predict the customer risk, thus investigating customer material of the low forecasted degrees of comparison can avoid deceiving policy effectively, and avoid service risk. Marisa .S.Viveros[1996]addresses the effectiveness of two data mining techniques in analyzing and retrieving unknown behavior patterns from gigabytes of data collected in the health insurance industry. J.-U. Kietz U. Reimer M. Staudt (1999)Mining Insurance Data at Swiss Life analysed to find out what the typical profiles of Swiss Life customers are with respect to the various insurance products.Mittal & Kamakura (2001) find the link between customer satisfaction and retention to be moderated by customer characteristics.kanwal garg(2008) find decision tree method for identifying customer behaviour of investment in life insurance sector. Patrick A Rivers(2010) examined some of the benefits and challenges of using data mining processes within the health-care arena. Data mining techniques can help insurance advisors to guide the potential customers and to map exact policy for proposal.The work in this paper is based on data collected from life insurance corporation of India and is aimed at predicting the policies of customer preference.

III.METHODOLOGY

Data mining methodology can often improve existing actuarial models by finding additional important variables, by identifying interactions, and by detecting nonlinear relationships.Insurance Market is purely based on customer penetration.Decision tree is a supervised data mining technique.It can be used to partition a large collection of data in to smaller sets by recursively applying two-way and /or multi-way spilts.Decision tree model is adopted to display relationship found in the insurance data set.The tree consists of zero or more internal nodes and one or more leaf nodes with each internal node being a decision node two or more child nodes.Using the data,the decision tree method generates a tree that consist of nodes that are rules.Each leaf node represents a classification or a decision.The training process that generates the tree is called induction.

ID3(Iterative Dichotomiser) decision tree algorithm is used for the analyses.ID3 uses information gain as its attribute selection measure.Let node N represent or hold the tuples ofm partition D.The attribute with the highest information gain is chosen as the splitting attribute for node N.This attribute minimizes the information needed to classify the tuples in the resulting partitions and reflects the least randomness or impurity in these partitions.Such an approach minimizes the expected number of tests needed to classify a given tuple.

The expected information needed to classify a tuple in D is given by
Info(D)=-Σ i=1 to m $p_i \log_2(p_{i)}$ Where pi is probability that an arbitary tuple in D belongs to class Ci and is estimated by |Ci,D|/|D| .Info(D) is known as entropy of D.Now we were to partition the tuples in D on some attribute A having v distinct values{a1,a2,…av}as observed from the training data.If A is discrete valued these values correspond directly to the v outcomes of a test on A. Attribute A can be used to split D into v partitions or subsets,{D1,D2,…Dv}where Dj contains those tuples in D that have outcome aj of A.
Info(D) is the expected information required to classify a tuple from D based on the partitioning by A. $Info_A(D)=\Sigma|Dj|/|D| * Info(Dj)$ The term |Dj|/|D| acts as the weight of the jth partition. Info(Dj) is the expected information required to classify a tuple from D based on the partitioning by A.
Information gain is defined as the difference between the orginal information requirement and new requirement.That is $Gain(A)=Info(D)-Info_A(D)$.The attribute A with the highest information gain (Gain(A)) is chosen as the spilitting attribute at node N Which leads to best classification.

IV.EXPERIMENTS AND DISCUSSION

A source data set was extracted from the Life insurance database.An initial trial dataset of some 10,000 records,covering a time period of 6-months June 2011 to December 2011.

The first task was to remove from the database those fields which were irrelevant to the task at hand.The second task is cleansing involved various transformations on the data (eg.birth dates transformed into ages).The third task involved collapsing the transaction oriented data that was supplied into a policy oriented dataset which is required for the types of analyses intended to be performed.
The segmentation of customers is done by considering demographic attributes of the customers.Variables considered for analysis are as Age,Gender: 0 Male,1 for female,Marital status,No of Kids- the customers of marital status married having possible values one or two or three.Service category :-customer may be a minor or major .Non-minor customer may be in still service or retired, Product type –Unit Linked –A,Traditional product-B, Plan types –Savings plans,Protection plans,Pension plans and Child Plans.

Decision tree is a popular classification method that results in a flow-chart like tree structure where each node denotes a test on an attribute value and each branch represents an outcome of the test.The tree leaves represent the classes.
.For the data set age values may be binned in to the following categories Unmarried in service,Unmarried not in service,Newly Married- in service-without kids,Married-in service-with Children,Married not in service,without children ,Married not in service with children. Their preference over the policy is transformed as A,B,C and D.In the above class Not in sevice refres to minor or jobless or retired from service.

Product type A refers to Unit-Linked and B refers to Traditional one.
Policy type refers to A For Savings,B Class refers to Protection plans,C refers to Pension Plans and D for Child plans.

The maximum number of levels in the tree in the tree was limited to four and the minimum number of records in a node was set to 2500,inorder to prevent the Decision Tree from becoming very complex.

The decision tree algorithm is to build a tree that has leaves that are as homogeneous as possible.The major step of the algorithm is to continue to divide leaves that are not homogenous in to leaves that are as homogenous as possible.
The decision tree alogorithm is given below.
There are 10000 samples.and four classes.

The frequency of classes as A=3000,B=2000,C=3000,D=2000
Information in the data due to uncertainty of outcome regarding the policy type

Each customer prefers to is given by

I=-(3000)/10000)log(3000/10000)-
(2000/10000)log(2000/10000)-
(3000)/10000)log(3000/10000)-
(2000/10000)log(2000/10000)
=(3/10)log(3/10)-(2/10)log(2/10)-(3/10)log(3/10)-
(2/10) log(2/10)
=0.5210895+0.4643856+0.5210895+0.4643856
=1.97

Each attribute in turn as a candidate to split the data set.

1.Attribute "Married" There are 7500 customers are married and 2500 that are not Value =yes has
A= 1250(1),B=1,C=3749(3),D=2500(2)
Value="no"has A=1249(2) ,B=1249(2),C=1,D=1

I(yes)=I(Y)=-(1250/7500)log(1250/7500)-
(1/7500)log(1/7500)-3749/7500)log(3749/7500)-
(2500/7500)log(2500/7500)
=0.43078+0.0124634549+0.5002+0.5288=0.472
=0.472

I(no)=I(N)=-(1249/2500)log(1249/2500)-
(1249/2500)log(1249/2500)-1/2500)log(1/2500)-
(1/2500)log(1/2500)
=0.500441112+0.500441112+0.0045+0.0045
=1.009

Total information of the two subtrees

=7500/10000 I(y)+2500/10000 I(n)
=0.75*0.472+0.25*1.009=0.606

2.Attribute "Service" There are 7000 customers are in Service and 3000 that are not.

Value="yes" has A=1, B= 2799,C=1400, D=2800
Value="no" has A=1499,B=1,C=1499,D=1

I(yes)=I(Y)=-(1/7000)log(1/7000)-
(2799/7000)log(2799/7000)-
(1400/7000)log(1400/7000)-
(2800/7000)log(2800/7000)

=0.0018+0.528+0.4643856+0.528=1.522

I(no)=I(N)=-(1499/3000)log(1499/3000)-
(1/3000)log(1/3000)-1499/3000)log(1499/3000)-
(1/3000)log(1/3000)
=0.500147346+ 0.00385+0.500147346+0.00385
=1.007

Total information of the two subtrees=7000/10000 I(y)+3000/10000 I(n)
=0.7*1.522+0.3*1.007=1.362

3.Attribute "Having Kids =Yes".There are 6000 having kids and 4000 that are not
Value="yes" has A=1,B= 1,C=2999 ,D=2999
Value ="no" has A=2000,B=1330,C=669,D=1

I(yes)=I(Y)=-(1/6000)log(1/6000)-
(1/6000)log(1/6000)-(2999/6000)log(2999/6000)-
(2999/6000)log(2999/6000)
=0.00209+-0.00209+0.500+0.500=1.004

I(no)=I(N)=-(2000/4000)log(2000/4000)-
(1330/4000)log(1330/4000)-
(669/4000)log(669/4000)-(1/4000)log(1/4000)
=0.5+0.52+0.431+0.0029=1.453

Total information of the two subtrees=6000/10000 I(y)+4000/10000 I(n)
=0.6*1.004+0.4*1.453=1.183

4.Attribute "Age subsection <40" There are 4000 are in the category <40 and 6000 that are not.

Value="yes"has A=1000,B=1000,C=1000,D=1000
Value="no"hasA=2000,B=1000,C=2000,
D=1000

I(yes)=I(Y)=-(1000/4000)log(1000/4000)-
(1000/4000)log(1000/4000)-(1000/4000)log()-
(1000/4000)log(1000/4000)
=0.5+0.5+0.5+0.5=2

I(no)=I(N)=-(2000/6000)log(2000/6000)-
(1000/6000)log(1000/6000)-
(2000/6000)log(2000/6000)-
(1000/6000)log(1000/6000)
=0.5283+0.431+0.528+0.431=1.9183

Total information of the two subtrees=4000/10000 I(y)+6000/10000 I(n)

=0.4 *2+0.6*1.9183=1.95

5.Attribute "Product type" There are 6000 are in the category A CLASS and 4000 that are not

Value="yes" has A=2000,B= 1000,C=2000 ,D=1000

Value="no"has A=2000,B=500,C=500,D=1000

I(yes)=I(Y)=-(2000/6000)log(2000/6000)-
(1000/6000)log(2000/6000)-
(2000/6000)log(2000/6000)-
(1000/6000)log(1000/6000)
=0.528+0.431+0.528+0.431
=1.918

I(no)=I(N)=-(2000/4000)log(2000/4000)-
(500/4000)log(500/4000)- (500/4000)log(500/4000)-
(1000/4000)log(1000/4000)

=0.5+0.375+0.375+0.5=1.75

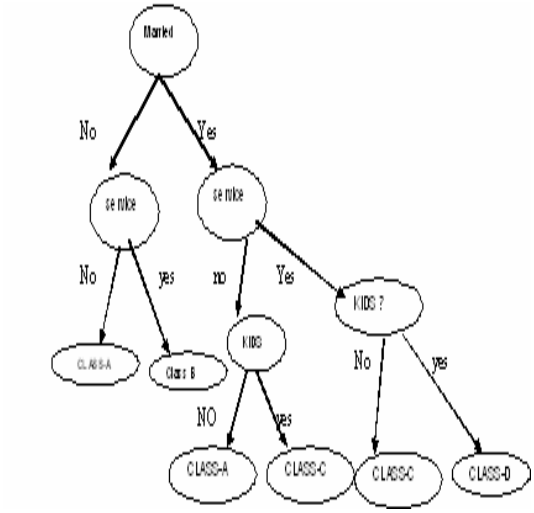
Total information of the two subtrees=6000/10000 I(y)+4000/10000 I(n)
=0.6*1.918+0.4*1.75=1.1508+0.7=1.850

Potential attribute	split	Information split	before	Information after split	Information gain
Married		1.97		0.606	1.364
Service		1.97		1.362	0.608
Kids		1.97		1.183	0.787
Age subsection		1.97		1.95	0.02
Product Type		1.97		1.850	0.12

Fig 4.1 Information Gain of five attributes

The above fig 4.1 gives Information gain determination with the information for the attributes before and after spilt. Hence the largest information gain is provided by the attribute "Married" and that is the attribute that is used for the spilt.
The attribute "Married" has the information gain and therefore becomes the splitting attribute at the root node of the decision tree.Branches are grown for each outcome of Married.The tuples partitioned accordingly.The resultant decision tree obtained with the spilt attribute is in figure 4.2.

fIg 4.2 Decision Tree for the dataset with the split attribute "Married"



The decision tree approach is widely used since it is efficient,can deal with both continous and categorical variables and generates understandable rules.The spilt variables used in building decision trees provide clear indication of which attributes are most important for classification.The above decision tree approach is able to deal with missing values in training data and can also tolerate some errors in the data.

V.CONCLUSION

The decision tree approach is widely used since it is efficient,can deal with both continous and categorical variables and generates understandable rules.ID3 decision tree algorithm is applied on the data set for predicting the customer preferences towards the policy preferences under the product type based on the spilt attribute.The adoption of supervised learning technique for prediction analysis in this paper used the demographic attributes of customer. The decision tree approach implemented in this paper clearly delineates the customer segment based on spilt attribute and contributes for retaining the profitable customers.
The segement of customers resulted based on spilt attributed can be utilized for cross selling and upselling of products.

References

1. Verhoef .C.,&Donkers .B.(2001).Predicting customer potential value an applicationin the insurance industry,Decision support systems,32,189-199

2. [Marisa S. Viveros,BM Research Division T. J. Watson Research Center]Applying Data Mining Techniques to a Health Insurance Information System, Proceedings of the 22nd

VLDB Conference Mumbai (Bombay), India, 1996

3. Kamakura, Wagner A. & Michel W. (2000). Factor analysis and missing data, Journal of Marketing Research, Vol.37, pp. 490–498.

4. Mittal, Vikas & Wagner A. K. (2001). Satisfaction, repurchase intent and repurchase behavior: Investigating the moderating effect of customer characteristics, Journal of Marketing Research, Vol.38, pp.131–142.

5. [Saundra Glover , Patrick A Rivers , Derek A Asoh , Crystal N Piper and Keva Murph](2010) Data mining for health executive decision support: an imperative with a daunting future., Health Services Management Research ,Volume 23, No. 1 , Pp. 42-46

6. Quinlan .J.R.,(1999,August) Simplifying decision trees,International Journal of Human-Computer studies,51(2),497-510.

7. Debahuti Mishra , Asit Kumar Das, Mausumi and Sashikala Mishram Predictive Data Mining: Promising Future and Applications-Int. J. of Computer and Communication Technology, Vol. 2, No. 1, 2010.

8. .Kim, Yong Seog, & Street, W. Nick (2004). An intelligent system for customer targeting: A datamining approach". Decision Support Systems, 37, pp.215–228.

9. Brusilovskiy, P. and Hernandez, R. (2001), Data Mining for a Sustainable Competitive Advantage. DM Review Magazine.

10. Seyed Mohammad Seyed Hosseini , Anahita Maleki, Mohammad Reza Gholamian (2010), Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty", Expert Systems with Applications, Vol 37, Issue 7, pp. 5259-5264.

11. E.W.T. Ngai , Li Xiu and D.C.K. Chau, (2009)Application of data mining techniques in customer relationship management: A literature review and classification, *Expert Systems with Applications*, Vol 36,Issue 2, Part 2 , pp 2592-2602.

AUTHORS PROFILE

S.Balaji received M.Phil computer science from ManonmaniamSundarnarUniversity,Tirunelveli, Tamilnadu in 2002 and MCA from Manonmaniam Sundarnar University in 1998.He is currently working as an Asst.Professor in MCA Department at Dhanraj Baid Jain College,Chennai,Tamilnadu,India.

Dr.S.K.Srivatsa received the Master degree in Electronics and Communication engineering from the Indian Institute of Science, in Bangalore. He received the Ph.D. degree in Electronics and Communication engineering from the Indian Institute of Science,Bangalore. Currently, he is a Senior professor at St Joseph's College of Engineering, Anna University, Chennai. His research interests include Digital Electronics, Networks and Communication, Graph Theory.