

Customer Segmentation with Clustering in Life Insurance Sector in India : A Review

ABSTRACT : Data mining is a technique for extracting useful information and discovering important associations between variables in a large data set or data warehouse. We study the application of several data mining approaches for knowledge discovery in the insurance industry in this paper. Data warehousing and data mining technologies have made a substantial contribution to practically every service sector, allowing organisations that provide services to clients to obtain competitive advantages. Data mining techniques are used to uncover hidden patterns in massive amounts of data stored in data warehouses. The integration of risk management with data mining will aid firms in more efficiently and effectively analysing risks and formulating risk mitigation and prevention strategies. Clustering algorithms can help marketing experts to achieve customer segmentation goals using K Means clustering.

KEYWORDS : Data Mining(DM), Insurance, K-Means, Clustering.



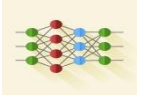

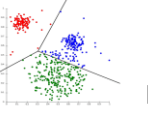
INTRODUCTION : Data Mining is a process of extracting previously unknown, valid, potential useful and hidden patterns from large data sets [1]. As the

amount of data recorded in CRM databases grows at an exponential rate. In the insurance industry, data mining can assist companies in gaining a competitive advantage. Companies may fully use data about consumers buying patterns and behaviour – as well as get a better understanding of their business – by using data mining tools to help prevent fraud, improve underwriting, and improve risk management. Risk management (RM) is becoming more widely acknowledged as a discipline that addresses both the positive and negative aspects of risk. Risk management is an important aspect of any company's strategic planning [2]. In the realm of safety, it is widely accepted that the only possible outcomes are negative ones, hence safety risk management focuses on preventing and mitigating harm. Risk Management is a tool used in the insurance industry to identify business opportunities.

DATA MINING : The practise of automatically uncovering usable information in big data repositories is known as data mining. Data mining techniques are used to sift through massive databases in search of new and

relevant patterns that might otherwise go unnoticed. They can also anticipate the outcome of a future observation, such as whether a freshly arrived customer will spend more than \$100 in a department shop [3]. The knowledge discovery in database (KDD) process relies heavily on data mining. Data selection, data cleaning, data transformation, pattern discovery (data mining), finding presentation, finding interpretation, and finding evaluation are all steps in the KDD process. The paper explores some data mining tasks, techniques, and applications [4].

Data Mining Is...

-  **Decision Tree**
-  **Nearest Neighbour Classification**
-  **Neural Networks**
-  **Rule Induction**
-  **K-means Clustering**

Data Mining is Not...

- Data warehousing
- SQL / Ad Hoc Queries / Reporting
- Software Agents
- Online Analytical Processing (OLAP)
- Data Visualization

LIFE INSURANCE IN INDIA : Life insurance is one of the fastest-growing sectors in India since 2000 as Government allowed Private players and FDI up to 26% and recently Cabinet approved a proposal to increase it to 49%. In 1955, mean risk per policy of Indian and foreign life insurers amounted respectively to ₹2,950 & ₹7,859 (worth ₹15 lakh & ₹41 lakh in 2017 prices). Life Insurance in India was nationalised by incorporating Life Insurance Corporation (LIC) in 1956. All private life insurance companies at that time were taken over by LIC. In 1993, the Government of India appointed RN Malhotra Committee to lay down a road map for privatisation of the life insurance sector [5]. In 2019, India's overall insurance penetration was 3.76 percent (life insurance 2.82 percent, non-life 0.94 percent), while the total insurance density was \$78 (life insurance density: \$58, non-life insurance: \$19).

TYPES OF LIFE INSURANCE IN INDIA :

- 1) Term insurance policies
- 2) Money-back policies
- 3) Whole life policies
- 4) Unit-linked investment policies (ULIP)
- 5) Pension policies








DATA MINING USED IN INSURANCE :

Data mining gives a number of insurance companies the capacity to forecast which claims are fraudulent, allowing them to better target their resources and recover large sums of money. This company can now anticipate which insurance claims are likely to be fraudulent thanks to data mining. Insurance is a unique type of product... Price optimization is a data mining technology that allows insurance companies to find out which groups of customers are more likely to accept a price increase and which are more likely to shop around for a new policy, according to Robert Hunter, CFA's director of insurance. The Comprehensive Loss Underwriting Exchange, or CLUE, and the less frequently used Automated Property Loss Underwriting System, or A-PLUS, are two databases that insurers use to track and transmit information about their customers.

ROLE OF DATA MINING IN INSURANCE INDUSTRY :

Data mining is a strong new technique that has a lot of promise for helping insurance companies focus on the most relevant information in the data they've acquired about their customers and potential customers behaviour [6]. Data mining help the insurance industry in identifying fraudulent claims and medical coverage, as well as predicting which customers will purchase new policies. Data mining is used in the insurance sector, and companies that have actually implemented it have a significant competitive edge.

Finding risk indicators that forecast profits, claims, and losses are some of the areas where data mining can be used in the insurance industry.

-  Analysis at the level of the customer
-  Analysis of marketing and sales
-  New product lines are being developed.
-  Reinsurance is a term that refers to the act of purchasing insurance
-  a financial evaluation
-  Estimating the provision for unpaid demands
-  Detecting and preventing fraud

PROPOSED SYSTEM : The proposed system's work focuses on creating the null hypothesis (H0) in order to see how different features and their combinations are related to the risk factor. Due to noisy behaviour, these

features may lack any relevancy and association among themselves, hence irrelevant features cannot be discarded in advance. When the policyholder's risk is not exposed to the insurance company, the system can also try to find out a solution to deal with asymmetric information. Different data mining strategies and approaches that are appropriate for features can be effective for reliable classification and clustering in order to understand some hidden intriguing data behaviour. As a result, this test is employed to test all-region on one side of the hypothesis. We only test the null hypothesis, and the alternative hypothesis (H1 or HA) is anything that isn't the null hypothesis.

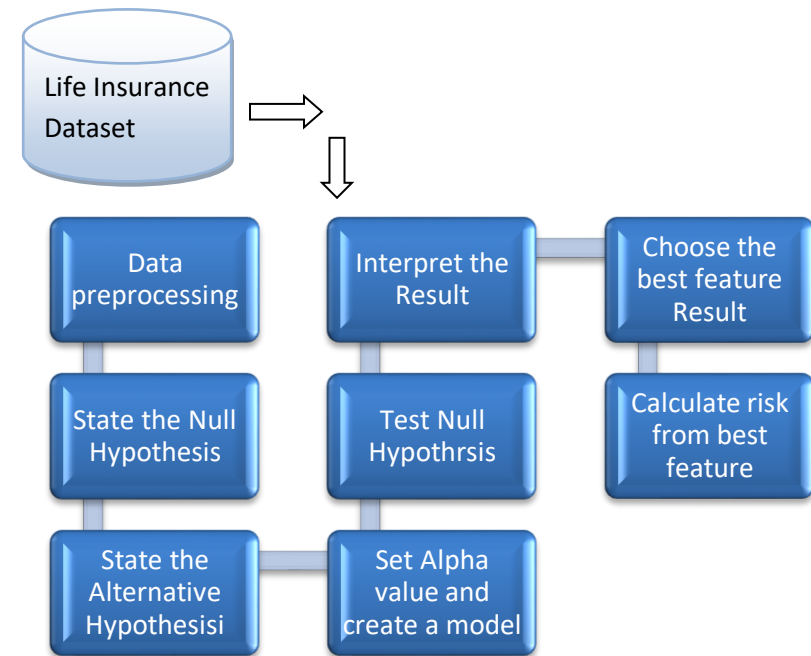


Figure: The step of proposed methodology

DATA MINING TECHNIQUES :

1. Naive Bayesian Classification Algorithm : Present actuarial models can typically be improved by using data mining techniques to locate extra important variables, identify interactions, and detect nonlinear relationships. Customer penetration is the sole determinant of the insurance market. Many machine learning and data mining approaches are based on Naive Bayes. Bayesian classification is a classification approach that can be used to classify large datasets. The hypothesis H of a naive Bayesian classifier is that the data tuple X belongs to a specific class C . $P(H/X)$ is the probability that hypothesis H holds given the evidence or observed data tuple X . The posterior probability of H

conditioned on X is $P(H/X)$. The Bayes theorem is important because it allows you to calculate the posterior probability, $P(H/X)$, given $P(X/H)$ and $P(X/H)$ (X) [7].

Bayes theorem is

$$P(H/X) = P(X/H)P(H)/P(X).$$

2. ASSOCIATION RULE : A difficulty defined in IAIS931 is the application of association rules to basket data type transactions. The mining of association rules is concerned with the discovery of intratransaction trends and is defined as follows:

Generate all associations so that the presence of a specified item(s) x in a transaction implies the presence of other item(s) y , given a database of

transactions where each transaction represents a set of items (e.g. services rendered). Given a support and confidence larger than a user-specified minimum support (s_i , $>$ and minimum confidence (c_{min}), the association rule $x \Rightarrow y$ will hold.

support is the number (or fraction) of the

transactions that contain a given item set.

confidence measures the frequency that items in a multi-item set are found together.

A data file, a relational table, or the output of a relational expression IAS941 could all be considered databases. The ability to handle huge datasets efficiently is one of this technique's benefits, and its execution time scales almost linearly with the size of the data. Throughout the course of our research, these qualities have proven to be true [8].

3. CLASSIFICATION : It is one of the most widely used strategies for developing models that can handle massive amounts of data. This technique is employed in the classifier training algorithm for company development. Decision tree, Bayesian classifier, neural network, Support vector machine, and other classifiers are used in classification techniques. Customer databases can be divided into

homogeneous categories, and classification divides data into pieces. On the insurance benchmark dataset, data mining classification methods are applied.

There are various classification types

- Supervised classification
- Unsupervised classification

4. CLUSTERING : It is used to identify objects that belong to the same class.

It's used to categorise customers depending on their behaviour. It can be used to segment customers and market to them more effectively.

Types of Clustering are

- Partitioning methods
- Hierarchical agglomerative methods
- Density based methods
- Grid based methods
- Model based methods

4.1 : FCLUST FUNCTION : On 1071 examples, we used the FCLUST function in R statistical programme version 3.2.2. The outcomes of the fuzzy clustering technique are examined in this section. Fuzzy clustering is a technique for

breaking M topics into K clusters by calculating and estimating the likelihood (factor) of each subject belonging to distinct groups. As a result, the generalised optimum value for customer clustering should be closely maintained. The Euclidean distance metric was utilised with a fuzzy parameter ranging from 1.1 to 4 to execute the algorithm in R. The number of clusters is determined by trial and error [9].

We began with 20 variables and ended up with 12 that were useful. By trial and error, the optimal number of clusters was determined to be two. Table 2 gives several examples of cases and their cluster membership values.

	Clus 1	Clus 2
Obj 1	0.36	0.64
Obj 2	0.31	0.69
Obj 3	0.52	0.48
Obj 4	0.44	0.56
Obj 5	0.66	0.34
Obj 6	0.39	0.61
Obj 7	0.69	0.31
Obj 8	0.43	0.57
Obj 9	0.64	0.36
Obj 10	0.63	0.37
Obj 11	0.35	0.65
Obj 12	0.34	0.66
Obj 13	0.42	0.58
Obj 14	0.33	0.67
Obj 15	0.29	0.71
Obj 16	0.59	0.41
Obj 17	0.43	0.57
Obj 18	0.48	0.52
Obj 19	0.66	0.34
Obj 20	0.69	0.31

Table 2 : The number of insured people in clusters who use the Fclust Program.

It is self-evident that the sum of membership values for all clusters for each object (case) is n1. i.e.

$$j_k = \sum_{j=1}^c 1, k = 1, \dots, n$$

The profiles of the two clusters are as follows, based on our analysis:

Profile of clients in 2- cluster analysis:

First cluster: Investment policy

- Around 36% of the cases are in this cluster.
- 60% of children under the age of nine.
- Women account for 51% of them.
- They are 89 percent single.
- 94 percent of the population is childless.
- They make up 47% of the population.
- 77 percent have a 30-year insurance duration.
- 56% of people pay on a monthly basis.
- 65% of the premium is less than 99000 Rials.
- 25% of the total capital (between 100 and 499 million Rials)
- Children account for 85 percent of those insured by policyholders.

As a result, the first cluster comprises of consumers who are students with no other means of paying the charge. This group is known as investment policy.

Second cluster: life security

- 64 percent of the cases are in this cluster
- 41 percent are between the ages of 20 and 29 years old
- 67 percent are married
- 51 percent do not have children
- 44 percent are self-employed
- 67 percent have 1 to 3 supplement coverage
- 70 percent have a 30-year insurance term

So, the second group of clients consists of married men who are self-employed, with 70% of them having a 30-year insurance duration. The total cost of the project is substantial. The name of this cluster is "life security."

	family	illness	weight	height	user	cover	fire	inv.life	annual	life
Clus 1	0.09	0.16	-0.52	-0.48	0.45	-0.39	0.01	-0.19	0.03	-0.07
Clus 2	-0.04	-0.10	0.38	0.35	-0.35	0.31	-0.03	0.16	-0.04	0.08
	The.premium	payment	insurance	Term.	job	Ratio	marry	children	gender	
Clus1	-0.01	0.04	0.02		-0.53	0.55	-0.38	-0.23	0.15	
Clus 2	-0.01	-0.05	0.01		0.41	-0.42	0.30	0.17	-0.14	
	Age	location	final.capital							
Clus1	-0.49	-0.03	0.00							
Clus 2	0.36	0.02	-0.01							

Table : Cluster centers in fuzzy clusters output

6. K-MEAN : K-Means is a clustering algorithm that uses a centroid or central point to connect each cluster (mean of points). During training, each item is assigned to the cluster with the nearest centroid, which is commonly calculated using Euclidean distance.

K-means algorithm

K-MEANS (\mathbf{D}, k, ϵ):

```

1  $t = 0$ 
2 Randomly initialize  $k$  centroids:  $\mu_1^t, \mu_2^t, \dots, \mu_k^t \in \mathbb{R}^d$ 
3 repeat
4    $t \leftarrow t + 1$ 
5    $C_j \leftarrow \emptyset$  for all  $j = 1, \dots, k$ 
   // Cluster Assignment Step
6   foreach  $\mathbf{x}_j \in \mathbf{D}$  do
7      $j^* \leftarrow \operatorname{argmin}_i \{ \|\mathbf{x}_j - \mu_i^t\|^2 \}$  // Assign  $\mathbf{x}_j$  to closest centroid
8      $C_{j^*} \leftarrow C_{j^*} \cup \{\mathbf{x}_j\}$ 
   // Centroid Update Step
9   foreach  $i = 1$  to  $k$  do
10     $\mu_i^t \leftarrow \frac{1}{|C_i|} \sum_{\mathbf{x}_j \in C_i} \mathbf{x}_j$ 
11 until  $\sum_{i=1}^k \|\mu_i^t - \mu_i^{t-1}\|^2 \leq \epsilon$ 

```

Figure : K-means Clustering Algorithm, adapted from [10]

6.1. EXPERIMENT :

We used the THE INSURANCE COMPANY (TIC) 2000 dataset for this study. The K-means algorithm will be used to perform this experiment [11].

Execution

We used the most 21 informative characteristics to the target attribute (caravan policy) in order to execute this experiment, and then we used the Clustering (K-

means) operator on the reduced dataset as indicated in the respected figure.

Clusters 0 and 2 are relatively close to each other, as shown in the table of centroids . As a result, they may share a number of qualities, as we will see in the next sections.



Figure : Rapidminer Process

Attribute Description

Customer Subtype
Customer main type
High level education
High status
Social class A
Rented house
Home owners
1 car
No car
National Health Service
Income < 30.000
Income 45-75.000
Average income
Purchasing power class
Contribution private third party insurance
Contribution car policies
Contribution fire policies
Contribution boat policies
Number of private third party insurance
Number of car policies
Number of boat policies

Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4
1-MOSTYPE	8.479	34.165	7.077	35.704	23.403
5-MOSHOO	2.444	7.933	2.020	8.413	5.192
16-MOPLHO	1.788	0.539	2.930	0.976	1.449
19-MBERHO	2.003	0.911	3.553	1.607	1.451
25-MSKA	1.635	0.639	3.162	1.425	1.226
30-MHHUUF	6.330	7.153	1.172	2.249	6.955
31-MHKOOP	2.682	1.857	7.838	6.756	2.055
32-MAUT1	6.313	5.693	6.778	6.013	5.284
34-MAUT0	1.944	2.547	0.988	1.624	3.277
35-MZFOND	6.135	7.483	4.634	6.320	6.934
37-MINKM30	2.692	3.942	1.108	1.946	3.972
39-MINK457	2.681	1.586	4.111	2.947	1.974
42-MINKGEA	3.694	2.921	4.934	3.955	3.078
43-MKOOKP	5.540	3.067	6.773	3.766	2.112
44-PWAPAR	0.682	0.711	0.854	0.710	0.951
47-PPERSAI	2.919	3.083	3.120	2.956	2.651
59-PBRAND	1.560	1.546	2.205	2.116	1.288
61-PPLEZIEI	0.019	0.012	0.027	0.020	0.015
65-AWAPAR	0.355	0.382	0.438	0.371	0.495
68-APERSAI	0.557	0.577	0.593	0.565	0.492
82-APLEZIEI	0.004	0.004	0.010	0.006	0.005

Figure : Centroid Table

Evaluation

The experiment, which used K-means on a smaller dataset, produced the following findings. Davies Bouldin's value is very little, indicating that the intra-distance (between points in the same cluster) is very short and the inter-distance (between clusters) is very large, indicating good clustering. For K=5, the Davies Bouldin value is 1.632, which is the smallest in our experiment.

Analysis

We showed the clusters, characteristics, and target class as a scatter plot for the most informative 10 attributes to read the clustering. The appendix section contains all of these scatter plots. We examined the plot that resulted and discovered the following intriguing findings:

Cluster A (C0) : In this cluster, singles, families with adults, seniors, retirees, and religious farmers are more likely to purchase a caravan policy. Due to the fact that a caravan is made of wood, consumers in this cluster frequently contribute 3-4 times in terms of fire insurance.

Cluster B (C1) : In this cluster Business Men and Retired, and Cruising Seniors, are less likely to have a caravan policy. This could be explained by the makeup of

these customers' families; for example, cruising elders would prefer to go on a cruise rather than camp out in a caravan, but their low purchasing power and average income could also be a factor.

Cluster C (C2) : Customers are well-off singles, middle-class families, or businessmen. According to their Intermediate to High Purchasing Power and Average Income, people in this category are more likely to possess a Caravan Policy.

Cluster D (C3): Seniors, Retired, and Religious Farmers make up this group. People who have contributed to four to seven automobile policies are more likely to get a caravan cover. This makes sense because any caravan must be towed by a vehicle.

Cluster E (C4) : Seniors, Retired, and Religious Farmers with the highest boat contribution have a larger likelihood of owning a caravan policy from other customers. Customers with two private third-party insurance contributions are also more likely to have a caravan coverage.

5. REGRESSION : It can be used to make forecasts. The link between one or more independent and dependent variables is modelled using regression analysis. In the insurance industry, more advanced procedures are required to forecast future values.

Types of regression includes

- Linear regression
- Non-linear regression
- Multi-variate linear regression
- Multi-variate non linear regression




6. SUMMARIZATION : With the use of visualisation tools, this report creation technique allows for better decision making for big volumes of client data. It will make business decision-making more functional. Data mining techniques can assist an organisation in addressing business challenges and making decisions, but choosing the right approaches is crucial.

Table1 : shows how data mining techniques will add value to insurance data.

Data Mining Techniques	Patterns
Classification	customer having similar characteristics Analysis of customer attrition in

	insurance sector Policy most likely to be used, most unlikely to be used Segments related to policy
Classification & Prediction	Predicting consumer behaviour Predicting the likelihood of success of policies Classifying the historical customer records Prediction of what type of policy most likely to be retained, most likely to be left Predicting insurance product behaviour and attitude Predicting the performance progress of segments throughout the performance period Prediction to find what factors will attract new avenues in Insurance sector Classify trends of movements through the organization for successful/unsuccessful customer historical records
Association	discovery of such association that promotes business technique
Summarization	provides summary information Various multidimensional summary reports Statistical summary information

DATA MINING TASKS IN INSURANCE INDUSTRY :

-  Acquiring new customers
-  Customer level analysis
-  Customer Segmentation

- ✚ Policy designing and policy selection
- ✚ Prediction
- ✚ Claims management
- ✚ Developing new product lines
- ✚ Underwriting and Policy management
- ✚ Risk management
- ✚ Reinsurance
- ✚ Fraud detection
- ✚ Trend analysis [12]

CONCLUSION : In the entire applications domain, data mining plays a critical role, particularly in the insurance sector. It emphasised the necessity and role of data mining techniques in the insurance industry for managing client data and gaining competitive advantage. It has the potential to improve the insurance business process. Data mining techniques are used in this arena for corporate decision making. This study examines how data mining approaches can help insurance companies cut costs, enhance revenues, recruit new consumers, retain existing customers, and develop new products. We may conclude that the insurance industry is growing at a rapid pace and employing numerous data mining approaches. When insurance companies utilise the right data mining tools, they may save a lot of money. Using innovative data mining techniques, the insurance industry's future lay in expanding product offerings and enhancing service standards.

REFERENCES :

- [1]-Khalili-Damghani, Kaveh, Farshid Abdi, and Shaghayegh Abolmakarem. "Solving customer insurance coverage recommendation problem using a two-stage clustering-classification model." *International Journal of Management Science and Engineering Management* 14.1 (2019): 9-19.
- [2]-Singh, D. and Kumar, P., 2012. Conceptual Mapping of Insurance Risk Management to Data Mining. *International Journal of Computer Applications*, 975, p.8887.
- [3]-Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. *Introduction to data mining*. Pearson Education India, 2016.
- [4]-Fu, Yongjian. "Data mining." *IEEE potentials* 16.4 (1997): 18-20.
- [5]- Wikipedia contributors. (2021, November 1). Life insurance in India. In *Wikipedia, The Free Encyclopedia*. Retrieved 15:04, December 25, 2021.
- [6]- A. B. Devalé and Dr. R. V. Kulkarni, "Application of data mining techniques in life insurance", *International journal of Data mining and Knowledge management Process*, Vol.2, No.12, pp. 31-40. July 2012
- [7]- S Balaji and S K Srivatsa. Article: Naive Bayes Classification Approach for Mining Life Insurance Databases for Effective Prediction of Customer Preferences over Life Insurance Products. *International Journal of Computer Applications* 51(3):22-26, August 2012. Full text available.
- [8]- [AIS93] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings @the ACM SIGMOD Conference on Management of Data*, Washington, D.C., May 1993.
- [9] - Iyzadparast S. M., Farahi A., Fathnejad F. and Pourteimour B. (2011) Using data mining techniques to predict the damage level of car insurance customers. *Journal of Information*

Processing and Management (former Information Science and Technology), 27(3):699-722.

[10] - Zaki, Mohammed J., and Wagner Meira Jr. Data Mining and Analysis: Fundamental Concepts and Algorithms. New York: Cambridge UP, 2014.

[11] - Namvar, M., Gholamian, M. R. and KhakAbi, S. (2010). A two phase clustering method for intelligent customer segmentation. In Intelligent Systems, Modelling and Simulation (ISMS), 2010 International Conference on (pp. 215-219). IEEE

[12]- Dilbag singh,Pradeep Kumar, "Conceptual Mapping of Insurance Risk Management to Data Mining", International Journal of Computer Applications, Vol. 39, No.2, pp.13-18 2012.