Customer Segmentation and Profiling for Life Insurance using K-Modes Clustering and Decision Tree Classifier

Shuzlina Abdul-Rahman, Nurin Faiqah Kamal Arifin, Mastura Hanafiah, Sofianita Mutalib
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA
Shah Alam, Malaysia

***Abstract*—Customer segmentation and profiling has become an important marketing strategy in most businesses as a preparation for better customer services as well as enhancing customer relationship management. This study presents the segmentation and classification technique for insurance industry via data mining approaches: K-Modes Clustering and Decision Tree Classifier. Data from an insurance company were gathered. Decision Tree Algorithm was applied for customer profile classification comparing two methods which are Entropy and Gini. K-Modes Clustering segmentized the customers into three prominent groups which are "Potential High-Value Customers", "Low Value Customers" and "Disinterested Customers". Decision Tree with Gini model with 10-fold cross validation was found as the best fit model with average accuracy of 81.30%. This segmentation would help marketing team of insurance company to strategize their marketing plans based on different group of customers by formulating different approaches to maximize customer values. Customers can receive customization of insurance plans which satisfy their necessity as well as better assistance or services from insurance companies.**

***Keywords*—Customer segmentation; customer profiling; decision tree; insurance domain; k-modes clustering**

I. Introduction

Insurance industry has been in the global market for decades and it is a critical contributor to a country's long term economic growth. Life insurers improve their policyholders' quality of life by pooling the risk of mortality, morbidity, and longevity among a wide number of people and returning the benefits of this pooling in the form of guaranteed payments [1]. In insurance industry, maintaining current customers is a challenge. Customer retention is more important than acquisition of new customers. It is said that 20% of the customers contribute more to the revenue of the company than the rest, as according to Pareto principle [2]. Despite the belief that clients are important for insurance organizations in gaining income and enhance their profitability, acquiring and retaining clients are serious issues faced by insurance firms [3]. It is not easy to obtain and influence new clients because when compared to the current clients, generally, new clients purchase 10% fewer than them, fewer involvement in the purchasing procedure as well as association with the seller [4]. Additionally, acquisition of new clients is more expensive compared to the maintenance of existing clients of the company [5]–[7]. Besides that, the likelihood of effectively selling a good or service to existing active clients is approximately 60-70 percent, while the likelihood is just 5-20 percent for potential clients, which made a greater likelihood of success in selling a good or service to existing clients compared to the potential ones [8]. It is also worthy to note that different clients contribute different amount of revenue to insurance companies, and so it is vital to handle clients based on their profitability due to uneven revenue generated by them [9].

Insurance companies are growing in numbers and the diversity of services offered, in which the clients have full control of their decisions [7]. It is thus important to have a good customer relationship management to retain the existing customers. To achieve that, insurance companies need to identify their target markets by segmenting the customers into groups. This allows them to choose whichever services that match their needs from any service providers. Customer segmentation helps business people to customize marketing plans, identify trends, plan product development, advertising campaigns and deliver relevant products, as well as personalizing messages of individuals for better communication with the intended groups [10]. Consumer sectioning is a great instrument in separating the consumers into various groups and perform analysis on their traits [3], and thus organizations are able to focus on clients in distinct features and determine the most valuable clients by sectioning the clients [9]. Clustering methods have been employed in many studies to segmentize customers [3], [9], [11]–[14], while classification via Decision Tree has also been widely used in past studies [15]–[17].

The following are the contributions of this paper:

- This research uses K-Modes Clustering and Decision Tree Classifier for customer segmentation and profiling for insurance domain.
- Marketing team of insurance company will be able to strategize their marketing based on different group of customers by formulating different strategies to maximize customer values.
- Customers can receive customization of insurance plans which satisfy their necessity as well as better assistance or services from insurance companies.

Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Shah Alam, Selangor, Malaysia.

The remaining of this paper is structured as follows: Section II discusses the related works on clustering and classification methods, while Section III describes the study's methodology. Section IV highlights the results and Section V provides the discussion and finally Section VI concludes the paper with future works.

II. Literature Review

A. Data Mining and Machine Learning

Investigation of unseen data and recognition of designs as well as affiliations that have valuable usages can be performed by information mining methods [18]. Organizations are able to pull out beneficial information from the data and obtain comprehension of their clients as well as their necessity through this information by implementing data-mining methods [7]. Data mining which is also part of knowledge discovery in database (KDD) involves the following process [19]: data selection, pre-processing, transformation, performing data mining algorithm, and data interpretation and evaluation. Data mining techniques like regression, classification, clustering, forecasting, association and visualization are also part of the classification framework in customer relationship management (CRM) [20].

While data mining extracting information from the vast amount of data, machine learning discovers algorithms that allows the machines to learn by itself without human intervention. Some examples of machine learning algorithms are Neural Networks, Decision Trees, Naïve Bayes, and Logistic Regression. K-Means, initiated by Mc. Queen in 1967 is the most popular and relevant clustering model [21]. Predictive classification models have been used to study customer purchasing behavior in past researches [13], [22], [23] in which classification models like K-Means and Decision Tree were commonly employed. This study explores these two models for segmenting and classifying customers.

B. Customer Segmentation via Clustering Methods

Past research shows that K-Means Clustering method has been widely used. K-Means Clustering was used to segment bank's customers whereby customers were grouped into five categories: potential growth customers, general customers, intermediate customers, senior customers and VIP customers [24]. Meanwhile, K-Means Clustering algorithm was also applied to segmentize private banking customers and the results showed three clusters named 'Core Value Customers', 'Financial Products Oriented Customers' and 'Deposit Oriented Customers' [25].

Khalili-Damghani et al. [9] employed K-Means Clustering for insurance customers segmentation. The results were three clusters labelled as 'profitable customer', 'potential profitable customers' and 'disinterested customers. Fuzzy C-Means clustering was used to cluster life insurance customers [26]. The results explained that two was the optimal number of clusters for the study which denoted as "investment" and "life security". In [3], a comparison of k- prototypes was conducted which combined K-means (for numerical element) and K-modes (for categorical element) algorithms, improved k-prototypes and SBAC (Similarity-Based Agglomerative Clustering (SBAC) for customer segmentation in auto insurance case study. The results showed that SBAC algorithm is more effective in clustering auto insurance customers with higher silhouette index value.

In another study by Qadadeh & Abdallah [27], K-Means Clustering and Self-Organizing Map (SOM) techniques were used to cluster insurance customers. The comparison was made between K-Means Clustering and the combination of SOM with K-Means Clustering which resulted in a better overall performance of the combined method with six clusters of customers. Further studies on SOM had been applied on imbalanced dataset for clustering categorical data in which Kohonen SOM (KSOM) algorithm was improved by focusing on the distance calculation amongst objects [35][36]. Another study on K-Means for clustering was done in [28] whereby K-Means algorithm was used to analyze the network traffic trend and type of traffic in campus network. The result showed that it was beneficial for managing or shaping the bandwidth usage and strengthens the security policy of the network.

K-Modes are generally the extended version of K-Means algorithm. The dissimilarity measure applied in K-Means algorithm is the reason that K-Means is unable to cluster categorical variables [29]. K-Modes clustering algorithm is introduced by Huang [30] by presenting a new measurement of dissimilarity to cluster categorical attributes [31]. While maintaining its proficiency, K-Modes clustering model eliminates the numeric data restriction. K-Modes removed the constraints imposed by K-Means through some adjustments including the usage of simple matching dissimilarity measure or hamming distance for categorical attributes and the replacement of means of cluster to the modes of cluster. The frequency-based approach is utilized by this model in updating the modes during clustering procedure to decrease the cost function which is estimated by calculating the standardized sum of within sum errors.

C. Rules Extration using Decision Tree Classifier

Clustering and classification techniques complement each other and are proved to perform well in segmenting customers. Clustering methods which are good at handling data without any labels have a setback of not being able to predict new and unknown data. On the other hand, classification methods are able to perform prediction to a set of unknown data but need to be trained by a set of labelled data. Decision Tree works in a way to guarantee the similarity of the sub-groups by splitting data points into two or more sub-categories [17]. A feature is represented by each node of the tree, and a value or a range of values for the feature that represents the node is portrayed by each edge aroused in a node [15]. The final output of the classification, known as class label, is stored in a leaf node. The comparatively straightforward process of Decision Tree makes it easy to understand and interpret, and the process that addresses a number of data intricacy that usually presents in the real data makes the method popular [32]. Hypotheses on each feature's own influence in the classification procedure are produced with the help of the decision rules uncovered on the pathways [15].

There are several applications that implement Decision Tree as classifiers. Clustering analysis using K-Medoid Clustering was performed on family farmers in Brazil, and

used Decision Tree aside from Support Vector Machine, Neural Network (Multilayer Perceptron) to identifying character that distinguish between those identified clusters [15]. Classifications using Support Vector Machine, Random Forest, Decision Tree, K-Nearest Neighbour, and Naïve Bayes were performed to predict the churning of credit card holders [16]. In a study by Ganjali and & Teimourpour [33] on life insurance customers, K-Means Clustering was used to group the customers based on their lifetime value. The researchers also performed association rules to the most valuable customer group as well as classification to predict position of new customers in each cluster. In [34], Decision Tree was used in job profiling analytics to select the most significant skillsets for each job position intelligently. It produced accuracy of 63.5% when used together with Capacity Utilization Rate. Decision Tree approach was used for classification process and the results showed that the model achieved 61.3% accuracy, 38.97 % classification error, 0.012% Kappa and 0.024% Correlation criteria.

Based on the previous research, K-Means Clustering and Decision Tree Classifier have been proven as the most popular clustering technique to group and classify customers across industries. Nevertheless, there is very limited study that perform and model categorical data which is proposed in this study. K-Means is only suitable for numerical data, whereas K-Modes is the extension of K-Means algorithm which can handle categorical data.

III. Methodology

This section presents the methodology of the study. Fig. 1 illustrates the four main steps which are detailed in the next subsections: 1) Data; 2) Variable Selection; 3) Model Development, and 4) Model Evaluation.

A. Data Preparation

The data used in this study was obtained from one of the life insurances companies in Malaysia. The total number of data was 37,181 records and it consisted of daily new business customers information including their demographic details and their policy information ranging from January 2018 until December 2019. Prior to conducting analysis, the data underwent a pre-processing phase including handling missing values, imputing outliers as well as transforming the variables. Missing values were imputed accordingly with blanks, while detected outliers were transformed. Data transformations methods include discretizing numerical variables via quantile-based approach, re-grouping of data in certain variables as well as changing data types. Discretization results in either conversion of some variables into categorical data, re-labelled to avoid redundancy or merged accordingly. Table A1 in Appendix shows the pre-processed variable description.

B. Variable Selection

Variable selection involves selecting attributes that provides meaningful insights towards targeting the right customers. Based on the data used, several variables were removed as they did not have impact in the analytics including 'Occupation Group', 'Distribution Channel', 'Insured (Self/Others)', 'Occupation Group', 'Payment Frequency', 'Premium Status', 'Race PO' and 'Sum Assured'. Additionally, business expert has suggested including some of the information regarding the policy purchased by the customers including duration of policy issuance, annual net premium, premium payment method, product type and policy status. Table I shows the final attributes selection.

C. Model Development

This study developed customer segmentation model using K-Modes and Decision Tree Classifier. Python language was used to perform the modelling for K-Modes Clustering and Decision Tree Classifiers. The first model, K-Modes was implemented with cost function in getting the minimize distance for the intra cluster distance. The number of clusters was set into k = 2, 3, 4 and 5. Then, the output for clustering is compared and evaluated in determining the best number of clusters. The optimal K value for K-Modes was determined by using Elbow Method. Fig. 2 shows that the elbow shape is detected when number of clusters suggested was 3 using cost function value and the precise value of the cost function is given in Table II. Therefore, K-Modes clustering with K=3 was chosen as the best number of cluster.

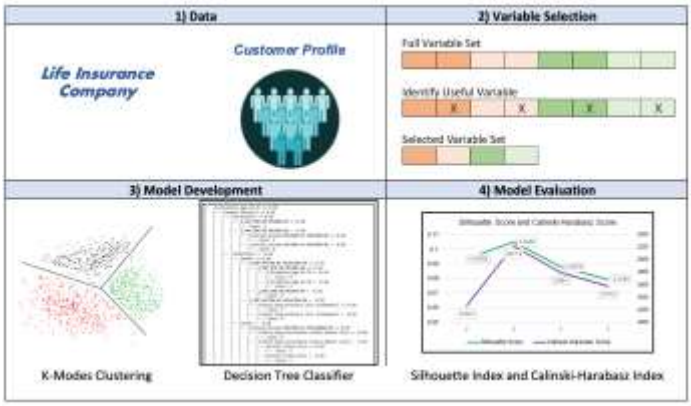

Fig. 1. Workflow Representation of the Methodology used in this Study.

TABLE I. Final Selected Attributes for Analysis

Index	Attribute	Attribute Type
1.	Annual Income PO	Ordinal
2.	Annual Net Premium (ANP)	Ordinal
3.	Client Type	Binary
4.	Duration of Issuance (Days)	Ordinal
5.	Gender PO	Binary
6.	Inception Age PO	Ordinal
7.	Location PO	Nominal
8.	Marital Status PO	Nominal
9.	Occupation Risk Class	Ordinal
10.	Payment Method	Nominal
11.	Active Policy (Y/N)	Binary
12.	Product Type	Nominal

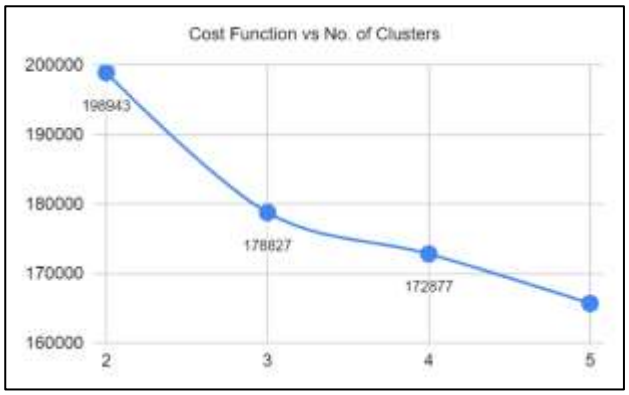

Fig. 2. Cost Function Plot.

The second model is a Decision Tree Classifier that was implemented with a built-in function of Python's Scikit Learn package. This function applies optimized CART (Classification and Regression Trees) in which it can perform well for binary classification as well as multi-class classification. The classifier was tuned by using the criterion function. This study experimented with two criteria of Decision Tree, 'Gini' and 'Entropy' with k-fold cross validation approach to achieve the best fit model. The number of labels was determined by the best number of clusters by K-modes; and in this study, 3 labels was defined for the classification task with rule extraction. The experiments were based on different number of clusters with evaluation of the cost function, thus the development time of clusters did not give any significant value, with 0.001 - 0.005 differences.

D. Model Evaluation

The clustering validation was done by using Silhouette Index score. The expectation of a good clustering is the shorter distance between each point in a cluster, the farther distance between clusters and a balanced proportion of data points among clusters. The equation of Silhouette Index is as shown in (1) [9]:

$$Silhouette = \frac{b(i)-a(i)}{\max\{a(i);b(i)\}} \quad (1)$$

where *a(i)* is the non-similarity between one object and other objects in the same cluster, and *b(i)* is the non-similarity between one object and other objects in the closest cluster. Evaluation using Calinski-Harabasz Index, CH was also performed to justify the performance of the clusters based on the formula shown in [9]:

$$\mathrm{CH}(q) = \frac{\frac{trace(B_q)}{(q-1)}}{\frac{trace(W_q)}{(n-q)}} \quad (2)$$

where *n* is number of records, *q* is number of clusters, W_q is intra-cluster scatter matrix, B_q is inter-cluster scatter matrix. The highest score portrays the best number of clusters for the dataset.

Meanwhile for the classification, the evaluation was performed using K-fold cross validation whereby the datasets were randomly divided into K equally sized subsets. The models were trained, and tested K times and the results were determined during each phase. The final accuracy of the models was measured as the average of all accuracies obtained in every iteration made. The formula for accuracy is as shown in (3) [15]:

$$\text{Accuracy } \% = \frac{c}{n} \times 100\% \quad (3)$$

where *c* is number of test samples classified correctly and *n* is total number of test samples.

IV. Results

A. Clustering Analysis

The results of distribution of each attribute in each cluster with K=3 is shown in Table A2 in Appendix. Based on the cluster analysis, it is shown that 51% (19,047) of the total observation falls under Category 0. This group has the highest percentage of young working customers with a relatively low annual income, and they aged in the range of 26 to 34 years old (37.7%) and earn MYR27,000.01 until MYR42,000.00 yearly (35.1%). The distribution of gender shows more than half of the customers are female and are married. Top residential location for customers in this group is Central Malaysia with 25.8, and slightly more than half of the customers in this group opt to pay low annual net premium which is in the range of MYR0 – MYR 1,800. Majority of the customers are new customers (88.9%) who purchased policies for the first time between of year 2018 and 2019. In addition, more than half of the customers belong to Class 1 of occupational risk which means that their occupations are having the least risk of exposure towards hazardous elements. For payment method, most of the customers use Auto Debit (80.5%) to pay their premium. The highest percentage of the policies' issuance days goes to '0 – 186 days' category (34.4%). Lastly, more than half of the customers purchase Ordinary Life (Endowment) products (68.7%) and their policies remain active (72.0%) at the end of year 2019. Hence, Cluster 0 is named as Low-Value Customers.

Cluster 1 makes up 27% (10,081) of the whole dataset. Customers are in young group aged between 10 to 25 years old (62%) and almost half of the customers have a low annual income from MYR0.00 – MYR27,000.00. Males are higher customers (68.8%) compared to female. Majority of the customers in this cluster are single (81.0%) and most of them reside in Northern Malaysia. In terms of the occupational risk class, more than of the customers belong to Class 1. Almost all customer in this group are new customers (90.6%). The annual net premium paid by the customers in this group are mostly between MYR1,800.01 – MYR 2,400.00 with 40.1% and they also prefer Auto Debit (75.7%) as the method to pay their premium. The distribution has the highest percentage for '187

– 334 days' duration policy (33.1%) and more than half of the customers purchase Investment-Linked (Whole Life) products. Lastly, the distribution of policy status is almost balance for this cluster with slightly higher percentage of inactive policies (55.3%). This means that this group of customers has higher chances to turn their policies inactive. Cluster 1 is labelled as Disinterested Customers.

On the other hand, Cluster 2 makes up of 22% (8,053) of the total observation. Majority of the customers in this cluster are older customers with more stable earnings since they have a high annual income. 49.3% of them age in the range of 42 – 76 years old and 61.8% of them have annual income in the range of MYR67,000.01 - MYR1,400,000.00. Moreover, more than half of the customers are male and 80% of the customers are married. Central Malaysia has the highest percentage for this cluster with 49.2%. This cluster also has the highest percentage of customers who belong to Class 1 hence their occupation is not very risky. Besides that, this cluster has a slightly higher percentage of existing customers (58.6%) compared to new customers.

Aligned with the range of annual income, this group of customers has the highest percentage of annual net premium in the range of MYR3,380.01 - MYR369,200.00 which is the highest category of annual net premium in this dataset with 51.5%. This cluster also has the highest percentage of those who issued their policy between 335 to 543 days with 35.5%. This group of customers prefers Credit Card the most with 58.4% as the medium to pay their premium to the insurer. For product type, the customers mainly purchase Investment-Linked (Whole Life) products with 79.5%. Finally, majority of the policies purchased are still active as of December 31st, 2019 with 84.3%. Cluster 2 is called as Potential High-Value Customers.

The cluster performance can be measured by evaluating intra-cluster performance and hence, we implemented Silhouette Index and Calinski-Harabasz Index. The results are shown in Table II. The cost function values are also included in the table for analysis purpose. Based on Table II, for Silhouette and Calinski-Harabasz Indexes, the scores need to be the highest to have the best cluster performance. In this study, it is shown that both scores are the highest when the number of clusters used are 3, as shown in Fig. 3. For the cost function, the value is decreasing when we add a greater number of clusters. However, the largest difference of the cost value is when the number of clusters is changed from 2 to 3 compared to the change from 3 to 4 clusters and 4 to 5 clusters which results in the elbow shape. Therefore, it is justified that K-Modes with 3 clusters has the best performance for this study.

B. Classification Analysis

The purpose of performing classification is to predict the characteristics of each class label by extracting the rules developed by Decision Tree. There was a total of 43 attributes including the target variable. We implemented K-Fold Cross Validation where it divides the dataset into K-folds and they have roughly the same size of samples. In this study, we performed experiments on both 'Gini' and 'Entropy' criteria for Decision Tree Classifier and the number of folds selected are 2, 5 and 10. We also set the maximum number of leaf nodes to 50 to ease the validation of the decision rules as this parameter enable the model to grow a tree in best-first decisions. Table III shows the outputs selected at random for all the experiments.

The performance evaluation of Decision Tree classification is done by comparing the accuracy of the models in each experiment. Since the experiments are implemented based on the k-folds cross validation method, the average accuracy for each model is compared. Referring to Table IV and Fig. 4, it is shown that the accuracy of the models increases as the larger value of K is used. It can be concluded that, Decision Tree classifier with Gini criterion and 10-fold cross validation is the best fit model for this dataset as it has the highest average accuracy compared to other models with 81.30%.

TABLE II. INTRA-CLUSTER PERFORMANCE EVALUATION

No. of Cluster	Cost Function	Silhouette Score	Calinski-Harabasz Score
2	198943.0	0.0935	1259.5
3	**178827.0**	**0.1049**	**2217.9**
4	172877.0	0.0878	1789.4
5	165769.0	0.0794	1574.6

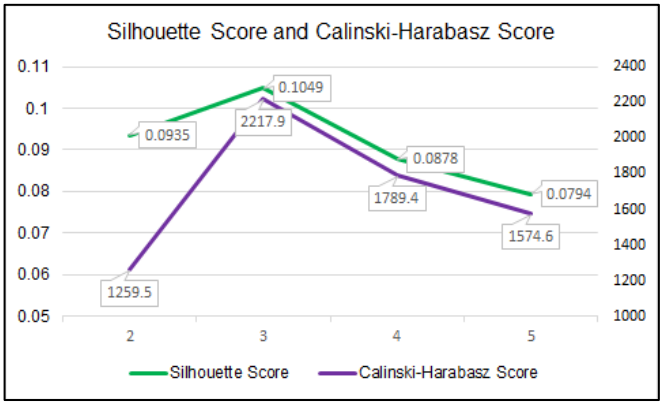

Fig. 3. Silhouette and Calinski-Harabasz Index for each cluster.

TABLE III. RULES OUTPUT GENERATED (RANDOM) FROM EXPERIMENTS

Exp.	Criteria	K	Rule 1	Rule 2	Rule 3
1	Entropy	2	If Marital Status is NOT Single, Payment Method is NOT Auto Debit, Annual Income is NOT MYR67000.01-MYR1400000.00 and Product Type is NOT Investment-Linked (Whole Life), hence, Cluster 0.	If Marital Status is NOT Single, Payment Method is NOT Auto Debit, Annual Income is MYR67000.01-MYR1400000.00 and Client Type is Existing Client, hence, Cluster 2.	If Marital Status is Single, Inception Age is NOT 10 – 25 Years Old, Policy is NOT Active, Product Type is NOT Investment-Linked (Whole Life) and Gender is Male, hence, Cluster 1.
2	Gini		If Inception Age is NOT 10 – 25 Years Old, Payment Method is NOT Auto Debit, Client Type is Existing Client, ANP is NOT MYR0.00-MYR1800.00, hence, Cluster 2.	If Inception Age is NOT 10 – 25 Years Old, Payment Method is NOT Auto Debit, Client Type is New Client, Gender is Male, Annual Income is NOT MYR67000.01-M1400000.00, Product Type is Investment-Linked (Whole Life) and Marital Status is Single, hence, Cluster 1.	If Inception Age is 10 – 25 Years Old, Gender is Female, Product Type is NOT Investment-Linked (Whole Life) and Annual Income is NOT MYR0.00 – MYR27000.00, hence, Cluster 0.
3	Entropy	5	If Marital Status is NOT Single, Payment Method is NOT Auto Debit, Annual Income is NOT MYR67000.01 - MYR1400000.00, Client Type is Existing Client and ANP is NOT MYR0.00-MYR1800.00, hence, Cluster 2.	If Marital Status is Single, Inception Age is NOT 10-25 Years Old, Product Type is NOT Investment-Linked (Whole Life) and Gender is Female, hence, Cluster 0.	If Marital Status is Single, Inception Age is 10-25 Years Old, Gender is Female, Product Type is Investment-Linked (Whole Life) and Annual Income is MYR0.00-MYR27000.00, hence, Cluster 1.
4	Gini		If Inception Age is NOT 10–25 Years Old, Payment Method is Auto Debit, Product Type is NOT Investment-Linked (Whole Life) and Gender is Female, hence, Cluster 0.	If Inception Age is NOT 10-25 Years Old, Payment Method is Auto Debit, Product Type is Investment-Linked (Whole Life), Gender is Female, ANP is MYR3380.01-MYR369200.00 and Client Type is Existing Client, hence, Cluster 2.	If Inception Age is 10-25 Years Old, Gender is Female, Product Type is Investment-Linked (Whole Life) and Annual Income is MYR0.00-MYR27000.00, hence, Cluster 1.
5	Entropy	10	If Marital Status is NOT Single, Payment Method is Auto Debit, Product Type is NOT Investment-Linked (Whole Life), Gender is Female and Client Type is New Client, hence, Cluster 0.	If Marital Status is NOT Single, Payment Method is Auto Debit, Product Type is Investment-Linked (Whole Life), Annual Income is NOT MYR67000.01-MYR1400000.00, Gender is Male, Policy is Active, Client Type is New Client and Inception Age is 43-76 Years Old, hence, Cluster 2.	If Marital Status is Single, Inception Age is 10-25 Years Old, Gender is Male, Product Type is Investment-Linked (Whole Life), Client Type is New Client, hence, Cluster 1.
6	Gini		If Inception Age is NOT 10-25 Years Old, Payment Method is Auto Debit, Product Type is NOT Investment-Linked (Whole Life), Gender is Male, Marital Status is Married and Client Type is Existing Client, hence, Cluster 2.	If Inception Age is 10-25 Years Old, Gender is Male, Issuance Duration is NOT 0-186 Days and Marital Status is Married, hence, Cluster 1.	If Inception Age is 10-25 Years Old, Gender is Male, Issuance Duration is 0-186 Days, Product Type is NOT Investment-Linked (Whole Life) and Policy is Active, hence, Cluster 0.

TABLE IV. MODEL EVALUATION RESULTS (K-FOLD CROSS VALIDATION)

K-Fold Cross Validation \ Criterion	**Entropy**	**Gini**
2-Fold Cross Validation	76.03%	77.81%
5-Fold Cross Validation	78.49%	80.19%
10-Fold Cross Validation	79.29%	**81.30%**

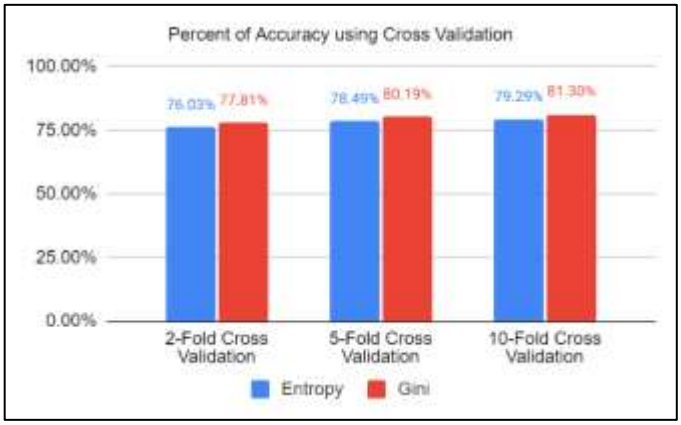

Fig. 4. The Accuracy of Classifiers for different K Fold.

V. DISCUSSION

Customer segmentation analysis is crucial for insurance companies to identify who are the profitable customers, how many percentages of them from the total population, to find more clients with similar profiles and how to manage less profitable clients [37]. Based on the results from this study, it can be concluded that customer segmentation can be achieved by using data mining techniques. Both clustering and classification methods are complementing the outcome of customer segmentation.

Once the customer segmentation is identified, there are some strategies that may be incorporated to cater each group of customers. For those customers who fall under 'Potential High-Value Customers', insurance company may want to focus more on this group to make the customers stay loyal to the company for a long period of time by providing good services to them. Since this group of customers contribute a lot to the company, the insurer may want to find more opportunities to sell more products to them based on their needs [9]. Insurer also needs to keep in touch with the customers from time to time to update their condition and keep on track of their well-being so that they feel comfortable with the company [9][38]. The insurance companies could build trusts relationship with the customers and this would increase cross-selling and up-selling [38]. Being

aware on customers' triggering events such as having a baby or buying a house could increase the product sales even more significantly, as according to the research done by [38]. Furthermore, insurer may provide a better customer experience by providing such a strategic and tactical focus based on the five key organizational process i.e., making strategic choices, creating value for customers, customer acquisition, customer retention, service quality and loyalty or rewards program, which can be achieved with a good CRM tool [39].

Nevertheless, the company must have strategies for those in group of 'Low-Value Customers' and 'Disinterested Customers'. For 'Low-Value Customers', although they are not the main contributors to the company, they are still the customers who are willing to take a chance in trusting the insurance company. Insurance companies may want to adopt customer centric approach to provide a superior customer experience [40]. This includes providing better customer services towards more customization and personalization by providing appropriate channels for communication to keep the customers informed, demanding and connected.

On the other hand, upon detecting the customers who fall into 'Disinterested Customers' group, insurance companies may be able to discuss and advise them to keep their policies in the event of customers cancelling the policies. In a study done by [39], it is evident that customers demand more on service quality, interaction management, contact programs, retention management, service strategy, customer satisfaction and customer loyalty, and this also could be achieved with a systematic CRM tool in place.

VI. Conclusion and Future Work

This study has presented the work on customer segmentation and profiling for insurance industry by using K-Modes clustering and Decision Tree Classifiers. The grouping of customers is made by analyzing the similarity of their characteristics and hence, able to determine the target customers. It is highly recommended for life insurance companies to segmentize their customers to enable them to offer suitable products or services in accordance with the needs of customers.

Future researchers may consider using a larger dataset with longer time periods to perform customer segmentation to have a more accurate result. If the data is too large, they may consider performing dimensional reduction technique such as Principle Component Analysis (PCA) to handle the data by transforming them into useful components. It is also suggested that future studies should use transactional details of the customers to monitor their behaviors and include more product categories such as Credit Life Insurance products as well as all rider products purchased by customers.

Further future work could also include result comparison with other classification models such as Random Forest, Naïve Bayes or even Artificial Neural Network (ANN). Also, computational complexity analysis could also be studied to analyze the learning efficiency and performance while implementing customer segmentation. The proposed approach in this study can also be applied in other industries like retail, hospitals, food chains, bookstores and so forth.

Acknowledgment

The authors would like to thank the Faculty of Computer and Mathematical Sciences, and the Research Management Centre (RMC), Universiti Teknologi MARA, Malaysia, for the support throughout this research.

References

[1] Cummins, M. Cragg, and B. Zhou, "The Social and Economic Contributions of the Life Insurance Industry," no. September, p. 39, 2018, [Online]. Available: https://brattlefiles.blob.core.windows.net/files/14446_life_insurance_industry_white_paper_final_2018.pdf.

[2] R. Srivastava, "Identification of customer clusters using RFM model: a case of diverse purchaser classification," Int. J. Information, Bus. Manag., vol. 9, no. 4, pp. 201–208, 2017.

[3] K. Zhuang, S. Wu, and X. Gao, "Auto insurance business analytics approach for customer segmentation using multiple mixed-type data clustering algorithms," Teh. Vjesn., vol. 25, no. 6, pp. 1783–1791, 2018.

[4] E. S. Levy, "Repeat Business is Online Retail" s Core Customer Metric," Thursday June 5th, eNewsletter. Core Cust. Metr., 2008.

[5] J. Griffin and M. W. Lowenstein, Customer winback: How to recapture lost customers--And keep them loyal. John Wiley & Sons, 2002.

[6] M. H. Hosseini, O. Mohammad MahmoudiMaymand, and M. Ahmadinejad, "Predicting the Bank Customer Switching Based on Data Mining Technique," Spectr. A J. Multidiscip. Res., vol. 2, no. 10, pp. 637–2278, 2013.

[7] F. Abdi, K. Khalili-Damghani, and S. Abolmakarem, "Solving customer insurance coverage sales plan problem using a multi-stage data mining approach," Kybernetes, 2018.

[8] M. Tarokh and K. Sharifian, "Applications of data mining in improving customer communication management," Iran. Ind. Manag. Stud. Q., vol. 6, no. 17, pp. 153–181, 2010.

[9] K. Khalili-Damghani, F. Abdi, and S. Abolmakarem, "Insurance customer segmentation using clustering approach," Int. J. Knowl. Eng. Data Min., vol. 4, no. 1, pp. 18–39, 2016.

[10] A. J. Christy, A. Umamakeswari, L. Priyatharsini, and A. Neyaa, "RFM ranking – An effective approach to customer segmentation," J. King Saud Univ. - Comput. Inf. Sci., 2018, doi: 10.1016/j.jksuci.2018.09.004.

[11] A. Ansari and A. Riasi, "Customer Clustering Using a Combination of Fuzzy C-Means and Genetic Algorithms," Int. J. Bus. Manag., vol. 11, no. 7, p. 59, 2016, doi: 10.5539/ijbm.v11n7p59.

[12] F. H. Bin Yusoff and N. L. A. B. Rosman, "A Case Study of Customers' Payment Behaviour Analytics on Paying Electricity with RFM Analysis and K-Means," in International Conference on Soft Computing in Data Science, 2019, pp. 40–55.

[13] N. M. N. Mathivanan, N. A. M. Ghani, and R. M. Janor, "Analysis of K-Means Clustering Algorithm: A Case Study Using Large Scale E-Commerce Products," in 2019 IEEE Conference on Big Data and Analytics (ICBDA), 2019, pp. 1–4.

[14] Y. Lu, A. Ioannou, I. Tussyadiah, and S. Li, "Segmenting travelers based on responses to nudging for information disclosure," e-Review Tour. Res., vol. 17, no. 3, pp. 394–406, 2019.

[15] C. Maione, D. R. Nelson, and R. M. Barbosa, "Research on social data by means of cluster analysis," Appl. Comput. Informatics, vol. 15, no. 2, pp. 153–162, 2019, doi: 10.1016/j.aci.2018.02.003.

[16] R. Rajamohamed and J. Manokaran, "Improved credit card churn prediction based on rough clustering and supervised learning techniques," Cluster Comput., vol. 21, no. 1, pp. 65–77, 2018.

[17] J. Asare-Frempong and M. Jayabalan, "Predicting customer response to bank direct telemarketing campaign," 2017 Int. Conf. Eng. Technol. Technopreneurship, ICE2T 2017, vol. 2017-Janua, no. December, pp. 1–4, 2017, doi: 10.1109/ICE2T.2017.8215961.

[18] Y.-H. Liang, "Integration of data mining technologies to analyze customer value for the automotive maintenance industry," Expert Syst. Appl., vol. 37, no. 12, pp. 7489–7496, 2010.

[19] K. Umamaheswari and S. Janakiraman, "Role of data mining in insurance industry," Int J Adv Comput Technol, vol. 3, pp. 961–966, 2014.

[20] E. W. T. Ngai, L. Xiu, and D. C. K. Chau, "Application of data mining techniques in customer relationship management: A literature review and classification," Expert Syst. Appl., vol. 36, no. 2, pp. 2592–2602, 2009.

[21] R. A. Soeini and K. V. Rodpysh, "Applying data mining to insurance customer churn management," Int. Proc. Comput. Sci. Inf. Technol., vol. 30, pp. 82–92, 2012.

[22] N. Isa, N. S. M. Yusof, and M. A. Ramlan, "The implementation of data mining techniques for sales analysis using daily sales data," Int. J. Adv. Trends Comput. Sci. Eng., vol. 8, no. 1.5 Special Issue, pp. 74–80, 2019, doi: 10.30534/ijatcse/2019/1681.52019.

[23] Y. B. Wah, N. H. Ismail, and S. Fong, "Predicting car purchase intent using data mining approach," in 2011 eighth international conference on fuzzy systems and knowledge discovery (FSKD), 2011, vol. 3, pp. 1994–1999.

[24] D. Dong, J. Zhang, and J. Ye, "Research on Customer Segmentation Method of Commercial Bank Based on Data Mining," no. Icidel, pp. 62–65, 2017, doi: 10.25236/icidel.2017.016.

[25] X. Yang, J. Chen, P. Hao, and Y. J. Wang, "Application of clustering for customer segmentation in private banking," in Seventh International Conference on Digital Image Processing (ICDIP 2015), 2015, vol. 9631, p. 96311Z.

[26] G. Jandaghi and Z. Moradpour, "Segmentation of Life Insurance Customers Based on their Profile Using Fuzzy Clustering," Int. Lett. Soc. Humanist. Sci., vol. 61, no. October 2015, pp. 17–24, 2015, doi: 10.18052/www.scipress.com/ilshs.61.17.

[27] W. Qadadeh and S. Abdallah, "Customers Segmentation in the Insurance Company (TIC) Dataset," Procedia Comput. Sci., vol. 144, pp. 277–290, 2018, doi: 10.1016/j.procs.2018.10.529.

[28] M. A. M. Ariffin, R. Ishak, S. A. Ahmad, and Z. Kasiran, "Network traffic profiling using data mining technique in campus environment," Int. J. Adv. Trends Comput. Sci. Eng., vol. 9, no. 1.3 Special Issue, pp. 422–428, 2020, doi: 10.30534/ijatcse/2020/6691.32020.

[29] S. S. Khan and S. Kant, "Computation of Initial Modes for K-modes Clustering Algorithm using Evidence Accumulation,", 2007, pp. 2784–2789.

[30] Z. Huang, "Clustering large data sets with mixed numeric and categorical values," in Proceedings of the 1st pacific-asia conference on knowledge discovery and data mining,(PAKDD), 1997, pp. 21–34.

[31] S. A. Sajidha, S. P. Chodnekar, and K. Desikan, "Initial seed selection for K-modes clustering–a distance and density based approach," J. King Saud Univ. Inf. Sci., 2018.

[32] H. A.Elsalamony, "Bank Direct Marketing Analysis of Data Mining Techniques," Int. J. Comput. Appl., vol. 85, no. 7, pp. 12–22, 2014, doi: 10.5120/14852-3218.

[33] M. Ganjali and B. Teimourpour, "Identify Valuable Customers of Taavon Insurance in Field of Life Insurance with Data Mining Approach Keyword :," vol. 4, 2016.

[34] E. A. Kamaru Zaman, A. F. Ahmad Kamal, A. Mohamed, A. Ahmad, and R. A. Z. Raja Mohd Zamri, "Staff employment platform (StEP) using job profiling analytics," Commun. Comput. Inf. Sci., vol. 937, no. September 2020, pp. 387–401, 2019, doi: 10.1007/978-981-13-3441-2_30.

[35] A. Ahmad, R. Yusoff, M. N. Ismail, and N. R. Rosli,, "Clustering the imbalanced datasets using modified Kohonen self-organizing map (KSOM)," 2017 Computing Conference, London, 2017, pp. 751-755. doi: 10.1109/SAI.2017.8252180.

[36] A. Ahmad, and R. Yusof, R, "A Modified Kohonen Self-Organizing Map (KSOM) Clustering for Four Categorical Data," Jurnal Teknologi, 2016, vol. 72 no. 1, pp. 1–6.

[37] C. Matiş, L. Ilieş, "Customer Relationship Management in the Insurance Industry," Procedia Economics and Finance, vol. 15, 2014, pp. 1138-1145.

[38] J. Godsall, A. Jain, K. Javanmardian, F. Nauck, S. Ray, and S. Yang, "Unlocking the next horizon of growth in the life insurance industry," McKinsey&Company, Sep 2017, https://www.mckinsey.com /~/media/McKinsey/Industries/Financial%20Services/Our%20Insights/ Unlocking%20the%20next%20horizon%20of%20growth%20in%20the %20life%20insurance%20industry/Unlocking-the-next-horizon-of-growth-in-the-life-insurance-industry.pdf

[39] E. A. Kumar, "Customer Relationship Management (CRM) Practices in Life Insurance Industry," Shanlax International Journal of Commerce 5 (4),2017, pp. 77-84.

[40] Y. Michaux, "Four ways insurance companies are improving their customer experience," IBM, 18 May 2021, https://www.ibm.com /blogs/services/2021/05/18/four-ways-insurance-companies-are-improving-their-customer-experience/

APPENDIX

TABLE AI PRE-PROCESSED VARIABLES DESCRIPTION

Attribute	Attribute Type	Description and Frequency of Data
Active Policy (Y/N)	Binary	Indicator whether the policy is still active 0: No (N) (32.8%) 1: Yes (Y) (67.2%)
Annual Income PO	Ordinal	Annual Income of Policy Owner (Customer) 0: MYR0.00-MYR27000.00 (25.0%) 1: MYR27000.01-MYR42000.00 (25.5%) 2: MYR42000.01-MYR67000.00 (24.6%) 3: MYR67000.01-MYR1400000.00 (25.0%)
Annual Net Premium (ANP)	Ordinal	Annual Net Premium (ANP) amount 0: MYR0.00-MYR1800.00 (38.0%) 1: MYR1800.01-MYR2400.00 (26.7%) 2: MYR2400.01-MYR3380.00 (10.3%) 3: MYR3380.01-MYR369200.00 (25.0%)
Client Type	Binary	Whether customer is an existing or new customerrecorded in the 2 years 0: Existing (20.9%) 1: New (79.1%)
Distribution Channel	Binary	Channel of which the policy is sold to the customers 0: Agency (39.4%) 1: Bancassurance (60.6%)

Duration of Issuance (Days)	Ordinal	Duration of days since the issuance date of thepolicy 0: 0-186 days (25.1%) 1: 187-334 days (25.0%) 2: 335-543 days (24.9%) 3: 544-729 days (25.0%)
Gender PO	Binary	Gender of Policy Owner (Customer) 0: Female (F) (50.7%) 1: Male (M) (49.3%)
Inception Age PO	Ordinal	Age of Policy Owner (Customer) at point ofpurchase of the insurance policy 0: 10-25 years old (25.3%) 1: 26-34 years old (27.6%) 2: 35-42 years old (23.0%) 3: 43-76 years old (24.1%)
Insured (Self/Others)	Binary	Whom the insurance policy covered 0: Others (22.5%) 1: Self (77.5%)
Location PO	Nominal	Customer's residential (in region) 0: Central Malaysia (28.3%) 1: East Coast Malaysia (5.4%) 2: East Malaysia (19.6%) 3: Northern Malaysia (26.4%) 4: Other Country (0.2%) 5: Southern Malaysia (20.1%)
Marital Status PO	Nominal	Marital Status of Policy Owner (Customer) 0: Divorced (1.3%) 1: Married (57.4%)2: Single (40.5%)3: Widowed (0.8%)
Occupation Group	Nominal	Occupation group of the customer 0: Housewife (1.6%) 1: Retiree (0.3%) 2: Self-Employed (1.1%) 3: Student (3.1%) 4: Unemployed (0.1%) 5: Worker (93.8%)
Occupation Risk Class	Ordinal	Risk class of the customer's occupation 0: Class 1 (Least hazardous) (73.4%) 1: Class 2 (Less hazardous) (14.3%) 2: Class 3 (Moderate hazardous) (5.3%) 3: Class 4 (Most hazardous) (7.0%)
Payment Frequency	Nominal	Frequency of the premium paid 0: Single which represented by 0 (0.9%) 1: Annually which represented by 1 (12.4%) 2: Semi-annually which represented by 2 (1.5%) 3: Quarterly which represented by 4 (3.2%) 4: Monthly which represented by 12 (82.0%)
Payment Method	Nominal	Premium's payment method 0: Cash / Cheque (8.0%) 1: Auto Debit (67.0%) 2: No Billing / Single (0.9%) 3: Credit Card (23.8%) 4: Advance Premium Payment (0.3%)
Premium Status	Nominal	Status of premium paying 0: Cancelled (5.5%) 1: Deceased) (0.0%) 2: Premium Holiday (4.1%) 3: Lapsed (22.6%) 4: Premium Paying (62.3%) 5: Single Premium (0.9%) 6: Surrender (4.5%) 7: Terminated (0.0%)

Product Type	Nominal	Type and sub-type of the product purchased bycustomers 0: Investment-Linked (Whole Life) (47.5%) 1: Ordinary Life (Endowment) (48.4%) 2: Ordinary Life (Hospital & Surgical) (3.0%)3: Ordinary Life (Whole Life) (1.2%)
Race PO	Nominal	Policy Owner's (Customer) Race 0: Chinese (52.5%) 1: Indian (21.2%) 2: Malay (18.6%) 3: Others (7.7%)
Sum Assured	Ordinal	Policy's sum assured amount. 0: MYR150.00-MYR13500.00 (25.1%) 1: MYR13500.01-MYR30000.00 (26.6%) 2: MYR30000.01-MYR100000.00 (34.8%) 3: MYR100000.01-MYR1800000.00 (13.5%)

TABLE AII DISTRIBUTION OF ATTRIBUTES IN CLUSTERS

Attribute	Cluster 0 (51%)	Cluster 1 (27%)	Cluster 2 (22%)
Active Policy (Y/N)	Inactive = 28.0% **Active = 72.0%**	**Inactive = 55.3%** Active = 44.7%	Inactive = 15.7% **Active = 84.3%**
Annual Income PO	RM0.00-MYR27000.00 =21.0% **RM27000.01-MYR42000.00 = 35.1%** RM42000.01-MYR67000.00 = 27.0% MYR67000.01-RM1400000.00 = 16.9%	**RM0.00-MYR27000.00 =47.9%** RM27000.01-MYR42000.00 = 19.9% RM42000.01-MYR67000.00 = 21.4% MYR67000.01-RM1400000.00 = 10.8%	RM0.00-MYR27000.00 =5.9% RM27000.01-MYR42000.00 = 9.5% RM42000.01-MYR67000.00 = 22.8% **RM67000.01- MYR1400000.00 = 61.8%**
AnnualNet Premium (ANP)	**RM0.00-MYR1800.00 =50.6%** RM1800.01-MYR2400.00 = 22.9% RM2400.01-MYR3380.00 =6.5% RM3380.01-MYR369200.00 = 20.0%	RM0.00-MYR1800.00 =35.9% **RM1800.01-MYR2400.00 = 40.1%** RM2400.01-MYR3380.00 =10.8% RM3380.01-MYR369200.00 = 13.2%	RM0.00-MYR1800.00 =11.0% RM1800.01-MYR2400.00 = 18.9% RM2400.01-MYR3380.00 =18.6% **RM3380.01-MYR369200.00 = 51.5%**
Client Type	Existing = 11.1% **New = 88.9%**	Existing = 9.4% **New = 90.6%**	**Existing = 58.6%** New = 41.4%
Duration of Issuance(Days)	**0-186 days = 34.4%** 187-334 days = 20.5% 335-543 days = 19.7% 544-729 days = 25.4%	0-186 days = 12.0% **187-334 days = 33.1%** 335-543 days = 26.3% 544-729 days = 28.6%	0-186 days = 19.7% 187-334 days = 25.5% **335-543 days= 35.5%** 544-729 days = 19.3%
Gender PO	**Female = 67.8%** Male = 32.2%	Female = 31.2% **Male = 68.8%**	Female = 34.6% **Male = 65.4%**
InceptionAge PO	10-25 years old = 15.1% **26-34 years old= 37.7%** 35-42 years old = 26.0% 43-76 years old = 21.3%	**10-25 years old = 62.3%** 26-34 years old = 16.5% 35-42 years old = 12.1% 43-76 years old = 9.1%	10-25 years old = 3.4% 26-34 years old = 17.8% 35-42 years old = 29.5% **43-76 years old = 49.3%**
LocationPO	**Central Malaysia = 25.8%**East Coast Malaysia = 6.2%East Malaysia = 23.7% Northern Malaysia = 23.7%Other Country = 0.0% Southern Malaysia = 20.6%	Central Malaysia = 16.4% East Coast Malaysia = 5.7%East Malaysia = 21.6% **Northern Malaysia = 36.1%** Other Country = 0.1% Southern Malaysia = 20.0%	**Central Malaysia = 49.2%**East Coast Malaysia = 3.4%East Malaysia = 7.3% Northern Malaysia = 20.4%Other Country = 0.6% Southern Malaysia = 19.1%
Marital Status PO	Single = 30.1% **Married = 67.9%** Divorced = 1.2% Widowed = 0.8%	**Single = 81.0%** Married = 17.5% Divorced = 0.9% Widowed = 0.6%	Single = 14.4% **Married = 82.7%** Divorced = 2.0% Widowed = 0.9%
OccupationRisk Class	**1 = 74.2%** 2 = 13.2% 3 = 4.5% 4 = 8.1%	**1 = 68.1%** 2 = 17.3% 3 = 7.2% 4 = 7.4%	**1 = 78.2%** 2 = 13.2% 3 = 4.8% 4 = 3.8%
PaymentMethod	C = 6.2% **D = 80.5%** N = 1.0%	C = 5.3% **D = 75.7%** N = 0.5%	C = 15.5% D = 24.2% N = 1.4%

Attribute	Cluster 0 (51%)	Cluster 1 (27%)	Cluster 2 (22%)
	R = 12.0% Y = 0.3%	R = 18.3% Y = 0.1%	**R = 58.4%** Y = 0.6%
ProductType	Investment-Linked (WholeLife) = 26.7% **Ordinary Life (Endowment) = 68.7%** Ordinary Life (Hospital &Surgical) = 3.5% Ordinary Life (Whole Life) = 1.1%	**Investment-Linked (WholeLife) = 61.2%** Ordinary Life (Endowment) = 36.4% Ordinary Life (Hospital &Surgical) = 1.4% Ordinary Life (Whole Life) = 1.0%	**Investment-Linked (WholeLife) = 79.5%** Ordinary Life (Endowment) = 15.3% Ordinary Life (Hospital &Surgical) = 3.7% Ordinary Life (Whole Life) = 1.5%

TABLE AIII LIST OF ABBREVIATION

ANN	Artificial Neural Network
ANP	Annual Net Premium
CART	Classification and Regression Trees
CRM	Customer Relationship Management
KDD	Knowledge Discovery in Databases
KSOM	Kohonen Self-Organizing Map
MYR	Malaysian Ringgit
PCA	Principle Component Analysis
PO	Policy Owner
SBAC	Similarity-Based Agglomerative Clustering
SOM	Self-Organizing Map