Journal of Finance and Accounting
2021; 9(6): 268-272
http://www.sciencepublishinggroup.com/j/jfa
doi: 10.11648/j.jfa.20210906.17
ISSN: 2330-7331 (Print); ISSN: 2330-7323 (Online)

Data-Driven Market Segmentation in Insurance Industry and Other Related Sectors

Chen Wen[1], Ke Gao[2], Yuanzhi Xiao[3]

[1]Olin Business School, Washington University in St. Louis, St. Louis, USA

[2]School of Economics, Beijing University, Beijing, China

[3]College of Liberal Arts, Texas A&M University, College Station, USA

Email address:
chen.wen@wustl.edu (Chen Wen), gkfly@126.com (Ke Gao), yuanzhixiao@yeah.net (Yuanzhi Xiao)

To cite this article:
Chen Wen, Ke Gao, Yuanzhi Xiao. Data-Driven Market Segmentation in Insurance Industry and Other Related Sectors. *Journal of Finance and Accounting*. Vol. 9, No. 6, 2021, pp. 268-272. doi: 10.11648/j.jfa.20210906.17

Received: November 2, 2021; **Accepted**: December 6, 2021; **Published**: December 7, 2021

Abstract: This paper talks about various approaches and models on customer segmentation in the insurance industry and other related sectors. In today's business world, especially the customer-centered industry, the most critical task is to find the right customers and serve the customers the way that most suits them. In this paper, we put our focus on the insurance industry for several considerations. One insurance company can possess hundreds of different policies, so it is crucial for policy issuers to find suitable policies for different customers. Considering the complexity and variability of different policies, insurance companies view customer segmentation as necessary and the key point for companies to compete well. Therefore, we select the insurance industry to study the effect of data-driven approaches on customer segmentation. In the first part, we discussed the need for a new approach to classify the customers and several advantages of the data-driven approach over the traditional method. In the second part of the paper, segmentation approaches such as K-means clustering, hybrid clustering, rule mining, and decision tree are discussed respectively about their processes and features. In the third part, we talked about the two current customer segmentation applications that are widely used today. We also talked about the segmentation systems in determining the risk of transmission of COVID-19. In the last part, we conclude the paper with the comparison of different approaches we discussed.

Keywords: Clustering algorithm, RFM Analysis, Decision Tree

1. Introduction

Market segmentation or customer segmentation is the core of any business model. It is an especially significant step in the insurance industry as every insurance company has hundreds of policies for applicants to select. Each policy has its characteristics and targeted group. However, some pairs or groups may have highly high similarities in insurance clauses or dividends systems in many such policies. Therefore, insurance companies must segment customers into different groups to determine a suitable policy respectively. In one research the author summarizes that segmentation can be done through different factors: demographic (Gender, age, religion), behavioral (consumption, spending), and psychographic (social circles, lifestyle). [1] The traditional method that insurance companies implement is based on pre-determined and experienced-based variables. Generally, it is an approach that relies on the factitious market segmentation, which has constrained segment space, limited data available, and bias towards certain variables. Due to these limitations, matching customers to policies by applying the traditional approach does not always yield optimized results.

Hence, the data-driven customer segmentation method is considered to be used in the insurance industry. One of the most significant advantages of the data-driven approach is that it does not rely on the intuition and experience of the business users. Instead, this approach provides a finer granularity of classification that is different from human thinking and traditional wisdom. The features of the applicants it captured are more detailed and broader. Nowadays, the data-driven market segmentation approach is prevalent in the strategic marketing area. This approach can be considered a member of data mining or machine learning. Many techniques, for example, soft computing, clustering, data visualization, and K-Means, are also used to develop the data-driven approach. Indeed, those

techniques gradually solve the problem and produce hundreds of variables to articulate the emerging market segments.

There is plenty of research on customer segmentation, not only in the insurance industry as well. Khajvand and Tarokh [2] investigated a framework to segment bank customers by a "score." The model gives each customer a score based on their deposit and transaction activities using the clustering algorithms. In the research paper [3], the author used the fuzzy analytic network process based on the RFM (Recency Frequency, Monetary value) model to implement the K-means clustering. The model recognized the most valuable and the riskiest customers for an auto insurance company. Researchers also could use data to avoid customer attrition. The researchers first gathered demographic data (gender, age, etc.) and studied their correlations. From there, they applied different selection methods to include the most critical variables into the model. [4] In another research the author compares two methods, decision tree, and logistic regression, to determine which model is better for predicting customers' behaviors. The result is that the decision tree is better than the logistics model with 70% accuracy. [5]

Nonetheless, further research on data-driven customer segmentation is expected by many insurance companies to produce better results. This paper mainly focuses on summarizing some current research on data-driven market segmentation in insurance companies.

2. Different Ways of Segmentation

In the research paper [6], the authors mainly focus on "presenting segmentation model for customers of Pasargad life insurance and then categorize them based on fuzzy clustering" [6] to illustrate techniques that can be applied to the data-driven customers segmentation approach. In this research paper, the authors aim to answer three questions. However, only two questions are mostly related to the broad topic of data-driven customer segmentation, and they are:

How can we segment life insurance customers using fuzzy clustering?

What are key variables in life insurance customer segmentation?

The data that the authors used are from the profile data of 1071 life insurance customers in Pasargadae life insurance company from March 2014 to October 2014. The first step that the authors considered is to determine the variables that can be used in the clustering method. By examining each variable offered in the contracts of Pasargadae life insurance company and getting advice from the insurance experts, authors decided on a group of 21 variables in total to use in the analysis, including age, gender, insurance term, number of supplementary coverages, etc. The main algorithm utilized by authors to construct the model is called FCM, Fuzzy C-Means Clustering, one of the most common approaches in clustering. FCM is used because dividing an item into a single cluster may result in some inaccuracy, which may be hard to generate an appropriate model. Moreover, "FCM clustering algorithm provides a method that enables data items belonging to two or more clusters, and most of this plan can be used in pattern recognition" [6]. The minimizing objective function is below:

$$J_{mf} = \sum_{j=1}^{N} \sum_{k=1}^{C} \mu_{jk}^{mf} \left\| x_j - C_k \right\|^2$$

Where $1 < mf < \infty$ is a real number, μ_{jk} is the membership degree of x_j to k-th cluster, x_j is the j-th sample and C_k is the center of the k-th cluster. The U matrix of FCM must satisfy $\sum_{j=1}^{C} \mu_{jk} = 1, \forall\, k = 1, \ldots, n$. Then we can get the segmentation by updating μ_{jk} and C_k via iterations:

$$\mu_{jk} = \frac{1}{\sum_{p=1}^{C} \frac{\left\| x_j - c_k \right\|^{\frac{2}{m-1}}}{\left\| x_j - c_p \right\|}} \text{ and } C_k = \frac{\sum_{j=1}^{N} \mu_{jk}^{mf} x_k}{\sum_{j=1}^{N} \mu_{jk}^{mf}}$$

By conducting the FCM technique in R, the authors find out that only 12 variables are influential, and "the optimal number of clusters was 2 which were named as investment and life safety" [6], which are a group that consists of students who are not able to pay for premiums and a group that consists self-employed married men with relatively long insurance period. From this research, two points are worth being aware of. Firstly, the fuzzy clustering method is a convenient and straightforward way to distinguish customers into specific groups, promoting the data-driven market segmentation approach. Secondly, using the clustering method is efficient for the decision-making process of the business user since it can analyze a large number of variables simultaneously.

In the research paper called Customer Segmentation in the Insurance Company (TIC) Dataset, written by Wafa Qadadeh and Sherief Abdallah, the authors mainly focus on using K-Means and SOM in the TIC 2000 dataset to interpret the data-driven customers' segmentation approach. This research paper wants to explore the central question: Can we discover meaningful clusters using different cluster analysis techniques applied on the high dimensional TIC CRM data [7]? K-Means clustering currently is not a famous clustering algorithm. It requires the analyst to specify the best number k of clusters to use. Furthermore, the authors applied the elbow method to find out the most appropriate number, k. Then, the author combined the K-Means method with SOM (Self- Organized Map), a kind of Neural Network that "converts multidimensional data into two-dimensional data representing the relationships between data objects" [7], to conduct data mining techniques.

Moreover, SOM can provide a clear visualization of the data. By executing only on the K-Means algorithm and the combined techniques of K-Means and SOM, the authors get very different results. The results are presented in the below table.

Table 1. *The Comparison Between K-means and Combined Method.*

Parameters	K-Means	SOM with K-Means
No. of Clusters	5	6
Execution Time	16 seconds	4 seconds
Davides Bouldin	1.632	0.699
Visualization	Difficult to visualize clusters	Clusters can be recognized easily

By taking a look at the result, a point is worth to be noticed. SOM with the K-Means technique is an effective way to do market segmentation. [7]

In another research paper, Behavior Segmentation based Micro-Segmentation Approach for Health Insurance Industry; the author discussed the necessity of customer segmentation for an insurance company. Identifying the customers that could bring the most profit to the company would be vital to the company's growth. On the other hand, to point out the customers with the most risk would be equivalently important.

Beyond the traditional segmentation method, such as demographic segmentation, the author introduced a new micro-segmentation that is segmented by customers' behaviors. [8] Such a segmentation method starts its work from the result brought by the traditional segmentation method. It uses machine learning technology and Artificial intelligence enhanced algorithms like RFM analysis to divide the customers into specific subgroups. [9] From there author explained what RFM analysis is

1. R stands for Recency, which is the days since the last claim on insurance.
2. F stands for Frequency, which represents the total number of claims per each policyholder.
3. M stands for Monetary to indicate the total claim charges for each policyholder.

From the RFM, the insurance company could identify each type of customer and provide specific strategies toward different customers. [8]

It takes many steps to perform the complete segmentation analysis. First of all, the researchers need to clean the dataset. To execute further research, they have to identify and remove all null values and possible outliers from the dataset. Second, they implemented the exploratory data analysis on the dataset. The process includes an overview of the dataset through summary statistics or data visualization methods. After the preparatory process, the researchers would use the traditional segmentation method, for example, through demographic information. Significant factors include age, gender, and education level. In the insurance industry, one would argue that each policyholder's age and health conditions are critical. After the demographic segmentation, the researchers would use the RFM analysis.

For each policyholder, the researchers first gather and group their data related to the claims. Based on the data of each factor of RFM, the model will rank the policyholders into different quartiles of R, F, and M.

Customer_ID	Recency	Frequency	Monetary	R_Quartile	F_Quartile	M_Quartile	RFMScore
849386.0	28	3	120702.47000	2	1	1	211
7930137.0	8	8	338279.21800	4	4	4	444
10293501.0	9	5	198634.62555	3	3	3	333
59223614.0	74	3	127614.44830	2	1	1	211
66950680.0	80	4	167888.76422	1	2	3	123

Figure 1. *The RFM score table.*

For each factor of RFM, the number 4 suggests the highest score of each RFM factor. For example, the second customer on the table who gets 4 in F-Quartile suggests that he/she belongs to the group with the most significant number of claims made.

Therefore, the RFM score of 444 suggests the type of customers that insurance providers should be cautious with: these customers made the claims most recent, most in quantity, most significant in monetary value.

Therefore, after implementing demographic segmentation, the RFM analysis of behavioral segmentation could provide more precise information about customer groups for the insurance company to make better strategic decisions.

While the model in this paper focuses on segmentation through the differences of the customers' claiming characteristics, some research papers provide a new way to use customer segmentation. In Multichannel segmentation in the after-sales stage in the insurance industry, the author suggests generational differences in communication channels with the insurances company. After conducting the research, the author concluded that users who communicate with a website live in regions with higher per capita income. Call centers users are more likely to live in regions with lower incomes. [10]

Using the data analysis on that study, the insurance company could allocate its marketing resources more efficiently. Another paper, Digital touchpoints and multichannel segmentation approach in the life insurance industry, shared a similar view on the value of customer segmentation through the channels they use. Questionnaires research this paper to determine how they search insurance information, and which purchase channel they prefer. Moreover, it also collected personal characteristics such as age and marital status. [11]. This survey could give the insurance company a better understanding of the channels the customers would use and how personal characteristics would change that. With this information, insurance companies could revise and improve their services.

Some approaches utilize multiple data mining algorisms. In the paper Hybrid soft computing approach based on clustering, rule mining and decision tree analysis for customer segmentation problem: The real case of customer-centric industries, authors proposed a hybrid model that is based on traditional clustering, rule mining, and decision tree analysis for customer segmentation problem. [12]

The model that came up by the researcher contains two

modules. In the first module, after the data cleansing, we could use k-means clustering to group the insurance company's customers based on their purchase behavior. The optimum number of clusters should be selected by the Davies-Bouldin index, which is based on a proportion of intra-cluster and inter-cluster distances [13]. After the clustering, the model would use the subset selection method to select the customer's features that might be statistically significant to predict the outcome. After the selection, each attribute of customers would be ranked with a score, highlighting the importance of each characteristic in customer segmentation. This paper illustrated an experiment with the dataset of customers of an insurance company in Iran, which eventually the model had Age with the highest score. [13]

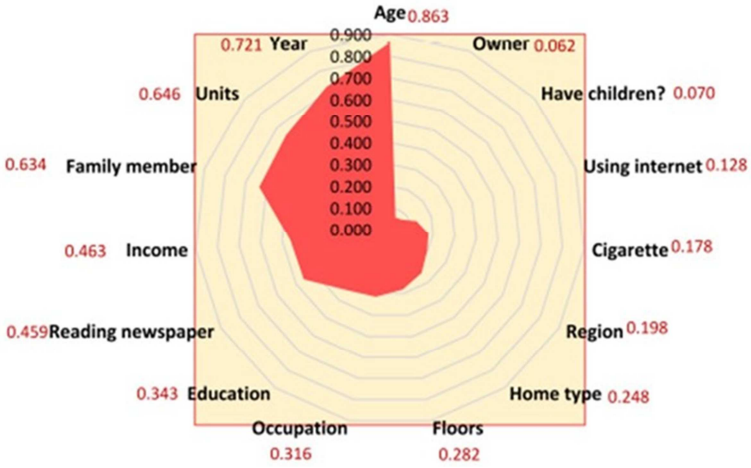

Figure 2. *Scores of Each Segmenting Factor.*

With the score being assigned to different attributes, the model furtherly extends with the decision tree method with IF rules to group the customers as High Value, Potential Valuable, and Low-Value customers. The researchers believe that the model could be more precise in customer segmentation with further improvement through machine learning technology.

3. Current Industry Practices and Applications

Thanks to the development of data science and AI technology, the practitioners have already built-up segmentation models that insurance companies could use directly.

Experian's Mosaic divides the whole U.S. consumers into 71 unique types and 19 overarching groups segments. This segmentation system utilizes more than 300 factors:

1. The customer character variable such as household income, baby number, age
2. Demographic variable such as city or countryside, zip-code
3. Customer behavior variable like happiness, culture to home or job

which provides a comprehensive vision to each group/ segmentation of customer.

Through Mosaic, Experian provides the insurance company with the most accurate and comprehensive profile of their customers and prospects by their similarities. The company can recommend the most suitable and even the most profitable policy to customers. In this way, the customer would become royal and satisfied with the service, and the company would reduce the customer churn rate with little effort. Instead of just making phone calls to individual customers, the data-driven way is cost-effective and can provide excellent care. In short, Mosaic offers a common customer language to define, measure, describe and engage target audiences through accurate segment definitions that enable more strategic and sophisticated conversations with consumers. Using Mosaic USA lifestyle segmentation, marketers can anticipate their best customers' behavior, attitudes, and preferences and reach them in the most effective traditional and digital channels with the best messages and digital channels with the right messages at the right time.

Another segmentation system is Esri Tapestry Segmentation. It contains 65 small groups and combines traditional demographic, psychographic, behavioral, and geographic market segmentations into one easy-to-use consumer segmentation system. Tapestry provides deep insight into the lifestyles and behaviors of diverse locations while offering a comprehensive view of the demographic populations within targeted markets. It involves segmentation by basic information such as customer locations and home addresses, including subfactors such as average income per capita by regions. It also has classification methods such as life mode, where customers are assigned to 12 groups based on their lifestyle and life stage. Tapestry is built using a unique combination of demographic and consumer data, paired with clustering methods and the latest data mining techniques. Esri's data scientists update the data annually, using the latest spatial boundaries, and always keeping data refreshed and accurate. The proprietary combination of rich datasets and geography results in multi-level 360-degree views of the consumer from national buying patterns down to community block groups.

A recent study utilizes a Tapestry segmentation system to predict a neighbor's COVID-19 infection rate and mortality rate. By correlating the COVID-19 data with the lifestyle segments from the Tapestry system, we could identify the patterns of transmission of the pandemic [14]. Furthermore, we could use the result to determine the factor that contributes the most to a region's health conditions. The findings could provide meaningful guidance for not only insurances companies but also authorities to implement improvement on the healthcare system more effectively. [15]

4. Conclusion

The models we have mentioned are all based on clustering techniques. From the difference in the result, we could learn that clustering alone is not sufficient for practitioners to reach a verdict on customer segmentation. The models that include multiple layers and steps, such as the hybrid model that incorporated clustering, rule-mining, and decision trees, performed better.

As insurance companies face an increasingly competitive environment, there will be a growing demand for advanced data analytics on customer segmentation. Such technology benefits not only the insurance providers but also the healthcare provider and policymakers. Segmentation could provide perspectives on their patients and constituents regarding their behaviors and status related to their health conditions. Using the information, the users could anticipate customer/constituents' future needs and behaviors and thus promote the general social welfare.

Acknowledgements

This work was supported by the Chinese Society for Technical and Vocational Education. “Research on the history and contribution of China's industrial development from the perspective of industrial economics” (SZ21D003) and “Research on increasing human capital investment and improving the contribution rate of vocational education”. (SZ21D004) Thanks for the support of Shandong Academy of Social Sciences. Collaborative innovation research on major theoretical and practical issues of social science planning in Shandong Province in 2020 “Research on the high quality development of economy promoted by Finance”. (20CCXJ24).

References

[1] R. Venkatesan, “Cluster Analysis for Segmentation,” University of Virginia Darden School, vol. 38, no. 91, pp. 92–98, 2007.

[2] Khajvand, M. and Tarokh, M. J. (2011). Estimating customer future value of different customer segments based on adapted RFM model in retail banking context. Procedia Computer Science, 3, 1327-1332.

[3] Ravasan, A. Z., & Mansouri, T. (2015). A Fuzzy ANP Based Weighted RFM Model for Customer Segmentation in Auto Insurance Sector. *International Journal of Information Systems in the Service Sector (IJISSS), 7* (2), 71-86. http://doi.org/10.4018/ijisss.2015040105.

[4] Goonetilleke, T. O. and Caldera, H. A. (2013) “Mining Life Insurance Data for Customer Attrition Analysis” *Journal of Industrial and Intelligent Information* 1 (1).

[5] Hassouna, M., Tarhini, A., Elyas, T. and AbouTrab, M. S., (2016). Customer Churn in Mobile Markets A Comparison of Techniques. arXiv preprint arXiv: 1607.07792.

[6] Jandaghi, G., & Moradpour, Z. (2015). Segmentation of life insurance customers based on their profile using fuzzy clustering. International Letters of Social and Humanistic Sciences, 61, 17-24. https://doi.org/10.18052/www.scipress.com/ILSHS.61.17.

[7] Wafa Qadadeh& Sherief Abdallah (2018). Customer Segmentation in the Insurance Company (TIC) Dataset. Published by Elsevier Ltd. https://creativecommons.org/licenses/by-nc-nd/4.0/.

[8] E. Y. L. Nandapala, K. P. N. Jayasena and R. M. K. T. Rathnayaka, "Behavior Segmentationbased Micro-Segmentation Approach for Health Insurance Industry," 2020 2nd International Conference on Advancements in Computing (ICAC), 2020, pp. 333-338, doi: 10.1109/ICAC51239.2020.9357282.

[9] 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.

[10] Dalla Pozza, I., Brochado, A., Texier, L. and Najar, D. (2018), "Multichannel segmentationin the after-sales stage in the insurance industry", International Journal of Bank Marketing, Vol. 36 No. 6, pp. 1055-1072. https://doi.org/10.1108/IJBM-11-2016-0174.

[11] Alt, M. A., Săplăcan, Z., Benedek, B. and Nagy, B. Z. (2021), "Digital touchpoints and multichannel segmentation approach in the life insurance industry", International Journal of Retail & Distribution Management, Vol. 49 No. 5, pp. 652-677. https://doi.org/10.1108/IJRDM-02-2020-0040.

[12] Kaveh Khalili-Damghani, Farshid Abdi, Shaghayegh Abolmakarem, “Hybrid soft computingapproach based on clustering, rule mining, and decision tree analysis for customer segmentation problem: Real case of customer-centric industries”, Applied Soft Computing, Volume 73, 2018, Pages 816-828, ISSN 1568-4946, https://doi.org/10.1016/j.asoc.2018.09.001.

[13] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-1, no. 2, pp. 224-227, April 1979, doi: 10.1109/TPAMI.1979.4766909.

[14] Ozdenerol, E., &Seboly, J. (2021). Lifestyle Effects on the Risk of Transmission of COVID-19 in the United States: Evaluation of Market Segmentation Systems. International journal of environmental research and public health, 18 (9), 4826. https://doi.org/10.3390/ijerph18094826.

[15] E. Engl and S. K. Sgaier, “Smarter micro-targeting to improve global health outcomes: scaling cluster segmentation on novel types of data for precision public health,” in 32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada, 2018.