Behavioural Cohort Analysis and Lifetime Value Stratification: AI-Driven Customer Segmentation Frameworks for Insurance Portfolio Management

Dr. Andrés Páez-Gaviria, Professor of Industrial Engineering, Universidad EIA (Colombia)

1. Introduction

It is increasingly important for insurance companies to adapt to the changing nature of consumers, who have diverse needs and demand coverage tailored to their particular lifestyle or specific event. Moreover, insurers operating in that sector of the market are facing intense competition that erodes their profit margins. To help tackle these challenges, customer segmentation has attracted growing scholarly and managerial interest. This paper presents some segments, which are developed using a cutting-edge AI method, namely artificial neural networks, from a dataset containing data on car insurers and policies. This helps fill a gap in the academic marketing literature because studies using AI neural networks for customer segmentation in the insurance context are scarce.

Driven in part by the internet, big data have begun to play an important role for insurance companies. Nevertheless, converting collected data into insight improves an insurer's decision-making for customer segmentation and, hence, provides a causal link with gaining a better understanding of their generalization of business practice. Moreover, customer segmentation for an insurer is an attractive way to attain their survival, not to mention gain a competitive advantage, because the traditional paradigm of insurance, which is based on traditional loss claims and customer profiles, is declining. Meanwhile, insurers strive to meet the needs and wants of customers and operate with thin profit margins.

1.1. Background and Significance

This chapter discusses the importance of innovative AI-driven customer segmentation methods in insurance, considering the recent developments and advancements in the

insurance market and AI technologies. The tremendous effect of technological advancements has led to major structural changes in consumer behavior, and this has, in turn, changed the dynamics of the insurance market. To address these changes in consumer behavior, companies must be abreast of these developments and offer personalized insurance advice. The delivery of good personalized offerings customized to customer-specific needs and preferences improves customer satisfaction and is a key driver for loyalty. Segmenting customers based on individual profiles and needs was relatively easier in traditional business models. It was based on simple demographics and stated payment and income.

Technological and customer behavior advancements have enhanced companies' ability to personalize offerings and have increased the benefits arising from such segmentation. AI is increasingly utilized to understand processes and patterns in large unstructured datasets. Multiple AI methodologies based on diverse techniques address customer segmentation. Often, studies focus on the quantitative value that such new segmentations may bear on performance metrics. This study is designed to highlight the different customer segmentation strategies used in the insurance industry and compare the traditionally used techniques with available AI technologies. It analyzes respective benefits, challenges, and provides a summary of the techniques in the industry. The application of AI technologies in this field is rather limited, and a significant portion of the industry is still stuck in ignorance. AI-driven segmentation techniques will allow for enhanced customer attractiveness and policy offerings. These offerings will be customized and personalized to a large extent, making procurement easier for customers. This will lead to substantial business growth in terms of both customer numbers and turnover.

1.2. Research Objectives

The primary objective of this study is to identify the ways AI-driven consumer segmentation is transforming insurance and the customer experience for policyholders. This objective will be achieved through pursuing the following specific research objectives: To explore how businesses are using AI-driven techniques to improve customer segmentation in insurance; To understand what is driving the development of AI for this purpose; To investigate how AI can improve accuracy and efficiency in segmenting customers; and To compare the effectiveness of a range of machine learning

algorithms at replicating human-performed customer segmentation in Nigerian communities. Despite rising interest in AI and the datafication of people and their possessions, there is a noticeable shortfall in practical effort seeking to experiment with insurance consumers in practice. We address these limitations, with the aim of taking a grounded approach, providing insurance companies with deep and actionable findings. In addition, this is the first study to compare the effectiveness of a range of machine learning algorithms for replicating human-performed customer segmentation. Improving such comparisons is an essential step in aggregating the lines of research on this topic, and as such, debates more distinction in machine learning methods in the business context. By identifying the most and least effective machine learning algorithms for the task, this study thus contributes to more robust and actionable insights for parties across the spectrum.

2. Theoretical Framework

Customer segmentation is a concept deeply rooted in academic and practical marketing theories. Among others, several traditional segmentation theories question the relevance of any further bundling of consumers into homogeneous groups. These theories and models created a strong foundation for such practical solutions as customer insights, targeting, proposition development, communication strategy, and thematic customer management. The introduction of technology into theories of customer segmentation and targeting was initially understood as a vehicle for refinement, extension, and optimization of existing market segmentation models, adding further dimensions or levels or providing real-time variations of existing segmentation models, either in single-sided or multi-layered approaches. More recently, the capabilities offered in segmentation and targeting have been extended beyond the use of extended multi-level cluster analysis models. With the development of machine learning, a new paradigm began to form which looks at customer segmentation and targeting from a completely different angle, i.e., a bottom-up, rather than top-down process. Such approaches pose that self-learning algorithms produce insights and predictions not explicitly programmed into existing theoretical models and may advance existing theories through the development of mega-segmentation models that account for individual-level interaction dynamics.

However, insurance companies seeking to use AI/ML in customer segmentation or any other process are faced with a seemingly unsolvable contradiction; on the one hand, they are advised to integrate insights and strategies from customer segmentation theories to ground AI/ML cluster formation models. On the other hand, advice given is that the concept more simply suffices in an environment that is becoming more and more volatile, and a step-by-step AI/ML model may not be effective within a few years' time because of the constant change. Moreover, one challenge in the insurance market is the fact that segmentation concepts are changing and expanding to consider, for instance, experience and engagement, product benefits and claims, or product purchase models.

2.1. Customer Segmentation in Insurance

Customer segmentation is the process of dividing a company's current and prospective customers or end users into subgroups or market segments based on the classification of certain traits. Consumer profiles help a business identify and target audiences that are more likely to invest in what they have to offer. In the insurance industry, profiling techniques are often used to divide consumers into groups based on certain personal characteristics. For insurance services, traditional segmentation criteria include demographic groups, psychographic profiles, and purchasing behavior. Many companies may use all three categories to build market segmentation profiles to target certain groups of people. Demographic segmentation practices have long been a strong foundation in insurance marketing. The typical consumer variables collected are generally those related to socio-economic status. Behavior-based segmentation in insurance identifies groups of customers or potential customers who do more or less of several activities. People in each different usage or habitual behavior or demographic traits group will have certain preferred action sequences. They thus can be targeted by the company to follow the desired action progression.

A recent growth in personalized insurance products is driven by the desire to serve certain market segments rather than the entire market in order to survive in increasingly competitive market surroundings. However, the continued rapid development of technology has sparked enthusiasm, particularly in the massive data handling and processing segment. The global insurance market is expected to grow significantly. Artificial Intelligence technologies, notably machine learning, are widely being embraced by the industry as a viable solution for a new segmentation approach that can

easily respond to an individual with its unique preferences and customization. In fact, new sophisticated algorithms for market segmentation have recently emerged for specific business sectors. AI-based solutions take insurance customer segmentation to new heights by combining data and behavior-centric perspectives. The essence of the practice is aimed at finding logical groups of relationships between variables and customers with similar characteristics, also referred to as market niches. This involves numerous distinct and cross-niche criteria that can be utilized to produce customer segmentation according to which all ad targeting, personalized approach, or service can be tailored.

2.2. Machine Learning in Customer Segmentation

Machine learning is often used as a digital transformer for customer segmentation. Based on the methods utilized, customer segmentation applied to insurance can provide valuable insights and establish a foundation to enhance mutual interaction between the client and the insurance company. Different algorithms such as decision trees, clustering, neural networks, or random forests are used. Machine learning algorithms offer automated and non-restrictive ways for data analysis and pattern recognition that can create biased and unbiased customer segmentation, compared to statistical or econometric methods. Nevertheless, data privacy and transparency approaches of the algorithms' functioning can be potential issues that companies have to address. Despite having sheer performance for the same datasets, not all machine learning algorithms are more suitable for a given business case. Moreover, some other issues regarding reliability and trustworthiness must be handled and thoroughly explored.

The enormous digital data agglomeration available right now has inspired companies to refunctionalize the way market strategy research is conducted. Developments in machine learning help companies better understand the market; simply put, machine learning can act as a digital transformer for customer segmentation. Customer segmentation is the analytical process of categorizing an individual customer, consumer group, or prospective consumer into specific types that describe the customer group's interaction with the insurer. Publicized customer behavior categorization allows better services and product adjustments, marketing strategy formulation, and value maximization of a given customer interaction. When considering the insurance market, customer segmentation helps companies identify profitable and non-profitable segments

and reduce promotional costs or churn rate strategies in the identified segment to provide new products or improve new segments.

3. Methodology

Research Design In line with contemporary research practice, the research has used a data-driven approach as a means to illustrate customer segmentation insights for insurance companies. In that regard, a research framework has been designed to create and implement the steps that are used to conduct the customer segmentation analysis and to address the research objectives. As a major stage in this research design, historical data had to be collected and processed to fit the purpose of the research. The research used a specific insurance market in a developing country as a population of interest, and a survey was utilized to collect the required data.

Data Collection Methods To achieve the research aims and objectives, data were collected through historical data analysis and surveys. Interviewing management personnel of the companies was an alternative option to be used in addition to the proposed methods, but because of the time and process constraints, no interviews were administered. The survey aimed to extract customer policyholder information with the permission of the insurance companies. This was because the targeted sample consists of policyholders of the insurance company population. For privacy concerns as well, the collection of the historical policyholder profile for each insurance company was not feasible when conducted outside the facade of MSEA. The data were collected from national insurance houses that operate in the country of Jordan. The target population was customers of motor insurance from the MSEA brokerage companies. Because of marketing, budget, and other limitations, the final questionnaire was distributed to 500 willing MSEA customers who are holders of car insurance policies. As for the guarantee of residency, the demographic information of the survey recipients was not uploaded in the dataset. Data collection duration was one month. Seventeen of the companies returned a response, whether electronic or hardcopy. Data from all companies were used to ensure a bias-free, generalizable, and homogeneous segmentation analysis.

3.1. Data Collection and Preprocessing

A variety of data types with multiple sources aid the accurate segmentation of customers and diverse classes of policyholders. Gathering the necessary data is one of the most challenging tasks in accurate and reliable customer segmentation. We used two

methods to get the data required for customer segmentation. Initially, we collected data by conducting a survey, and then we performed an analysis on existing datasets. One of the key challenges of data collection is to manage flexibility in choosing an appropriate source for collecting the data. Data must match the policyholders' profiles regarding the insurance market and user retention orientation. However, the data collected directly from the users in the field survey might not have sufficient validation and quality assurance processes. Moreover, in a closed environment like the insurance industry, the data owner and the surveyor might have potential issues related to privacy concerns or incomplete or fake data. High-quality novel data sources have sensitive industry-specific characteristics, and they pose serious ethical, legal, and technical challenges. Another major challenge is to have a structure of data that is understandable by machine learning methods for customer segmentation.

For many data sources, the data needed for customer segmentation is often grossly incompatible. This incompatibility between data is often characterized by the differences in the structure, granularity, and quantity. Under these circumstances, combining these data sources is not straightforward. Simply merging data of various granularities and sizes may create noisy data, which can affect model learning and customer segmentation results. Therefore, after obtaining the data, we further process it by using a data preprocessing technique to make it of suitable quality for machine learning analysis. The preprocessing tasks that we have chosen to perform are data cleaning, data normalization, and data transformation. During the data normalization process, we will transform the values of numerical variables to be between 0 and 1. The purpose of the data transformation is to change the scale of numerical data and test the correlation matrix. The information about the distribution of data can be valuable in data preprocessing to profile its characteristics for the unsupervised task of choosing an appropriate segmentation algorithm. In this work, we analyze the data distribution of each variable and choose the appropriate algorithm that can work with the dataset's distribution characteristics.

3.2. Algorithm Selection and Implementation

Considering the different types of analytical borders in combination with the strong separation of data instances validation with silhouette coefficients, three analytical distances could have been relevant: Euclidean, geodetic, and Bombardier-Jones. These

scores recommended classifying policyholders into four clusters and were useful for defining potential policyholders' values. Despite these preliminary insights, it is noted that the previous exercise clarified that other models could exist and might better represent the data. Hence, a more advanced approach combining the use of different machine learning and deep learning models is tested. As four parts seem to classify policyholders well, hard clustering models, i.e., K-means and BLAST, as well as standard GMM and DBSCAN, are tested. Hyperparameter selection and implementation for the standard K-means model are used. The use of random initial seeds in K-means can lead to different solutions, so the model is run several times to get the most balanced outcome. In the other versions of K-means, initial seeds are set only once. Moreover, since the three other algorithms contain some form of automatic tuning, multiple algorithms will test them relatively quickly to understand their best fit. For some models, the computational load is too high to estimate the outcome. If possible, through clustering performance metrics, the feasibility of the alternative algorithms will be verified, i.e., homogeneity, completeness, and silhouette. These metrics reveal the compactness of the clusters. For that performance measure, results closer to 1 are better. Implementability measures, especially Euclidean homogeneity and completeness clustering, have strong effects on data performance. An ensemble between these measures is then tested to ensure that the ability to capture clustering better is reflected in the silhouette.

4. Results and Findings

Based on the conceptual and empirical research results, as well as the guidelines of scientific methodology, the results of the research and analysis of the data and information were obtained. A customer segmentation for practical business work applies a variety of analysis algorithms. In the study, the k-prototype, k-means, and hierarchical segmentation of customers by age, gender, sickness, coverage level, service claims, experiences, and satisfaction are the primary activity analysis algorithms. Each of the algorithms clusters customers in a different solution, and its computation of performance success is determined by the use of performance metrics indicators. The results of the application of the performance metrics indicators of the algorithms show that based on the value of silhouette, which is 0.615, the calculation results of TSS is 645.189, and the value of BSS is 1003.869 for the k-prototype algorithm. The k-means clustering with a number of clusters equal to 5, and hierarchical clustering with a

number of clusters equal to 5, produce the best results for the process of customer segmentation.

Restrictions on insurance services using the application of the k-prototype algorithm, k-means, and hierarchical clustering to data claims in the insurance segment are: (1) the assumptions known a priori are not appropriate or relevant to the clustering results; (2) no membership assignment during the LATA in each segment; (3) insurance relevance due to some of the results of each clustering segment, so that the wave of membership produced by the clustering segment results does not satisfy the fuzzy, hard, and crisp criteria. The implications of limitations in this research are that the grouping process will produce several eligibility criteria for proposed policy status that guarantee the wave of membership, plus the production of various sub-criteria granting rights to the segments based on claims data ex-ante of the insurance in other scopes or insurance companies in Indonesia. Based on the analysis of the obtained pattern segmentation of customer turnover, it was found that they could be categorized into seven specific classifications of customer turnover patterns.

4.1. Performance Evaluation of Algorithms

One of the main advantages of the selected machine learning algorithms is their capability to work with non-linear dependencies in the data. To identify machine learning models with the best customer segmentation capabilities, the performance evaluation metrics are applied. We evaluate classification effectiveness based on the following metrics: accuracy (accuracy of the overall classification result), precision (the number of relevant elements selected), recall (the number of relevant elements that are selected by algorithms), and the F1 score (the harmonic mean between precision and recall). These metrics are chosen to demonstrate algorithm performance to service providers. In each table, the best algorithm performance metrics are highlighted for all evaluation measures.

Cross-validation produces comparable results. We select the most successful machine learning models for further analysis and practical application as the predictive engines for customer segmentation in the insurance industry. The practical implications of this research include the selection of model algorithms that are formulated as computing engines when considering the acquisition of technological solutions in real insurance company scenarios. Furthermore, as the research is based on real data, we suggest

comparing the predictive models of our study with other models in customer segmentation contexts. Identifying a leading algorithm or ensemble of multiple algorithms provides valuable information for industry experts. The real value of the proposed division lies in choosing the best technologies for commercial use. We choose the best algorithms and integrate them into existing infrastructures in the insurance company.

5. Discussion and Implications

The magnitude of theoretical improvement, research activity resulting from region-wise initiatives, applications in industry, etc. do not follow automatically by drawing numbers from a normal distribution. This pattern, referred to as the long-tail distribution, is commonly seen in real-world markets and systems, and is consistent with a range of theoretical frameworks. If one translates these results into what it means for insurance, and in particular for customer segmentation, the practical applications are especially relevant. There seems to be evidence that certain policyholders are shifting between regions of the distribution, and that these shifts appear linked to several underlying determinants.

Practically speaking, this means that insurance companies can be more effective in reaching out to those customers that could benefit from more active engagement through more tailored offerings, marketing, and pricing, and thus by doing so ensure that over time they have a better meeting of both underwriting and customer needs. This paper has outlined several ways in which the practical implications of our findings can be integrated into an insurance company's decision-making practice. As more and more insurance firms have or are dedicating substantial resources and time to creating a differentiated approach to offering, pricing, and communication, there is evidence that they are looking for this level of detailed analysis. However, this paper focuses on this as a 'next steps' approach by utilizing advanced machine learning in describing and segmenting their customer base. Given the practical limitation of time and general capabilities for machine learning, understanding the key areas that tend to come out first as a focus for different segments is important, but also an increasing area for future market insight research alongside decision behavioral modeling research. Overall, the application of machine learning in customer segmentation has implications and impact for rational and behavioral analysis in insurance, and has a much wider industry

perspective. Firstly, large amounts of customer-level data suggest that along with pricing and underwriting innovations, the need for more effective segmentation is becoming an increasingly important issue. While the insurance industry is confronted with a number of technological developments that could be seen as challenges, it is the distinction and discrimination of the insurance business model, including risk accumulation and spreading, that will continue to require the ultimate 'human judgment'. Besides, if we tie this to further developments such as the current and developing dynamics of insurance company strategies, these have the potential to impact the policyholder-distributor and reinsurer networks, while on the one hand retaining share and operating in newly adapted market segments, and on the other unlocking various competitive technologies. We believe that the work described in this paper provides a useful platform and synergy for industry practitioners adapting to these many interconnected challenges. Our future work will examine this scenario from a market perspective with a more thorough grounding in modeling and decision behavioral science.

5.1. Interpreting the Results

Our task is to interpret the segmentation outcomes derived from data-driven algorithms and compare those with the segmentations established in theory. This is to assess both the relevance and obtainability of a supposed golden rule when using data-driven AI technology to uncover and support insurance customer needs. The gap analysis between our results and the theoretical segments shows both theoretically interesting and practically relevant implications. For insurance practitioners, the segmentation might be informative to adapt their targeting strategies. It not only reveals practical findings with respect to customer profiles but also provides insight into consumer behavior, as association rules are made based on historical premium data available before these customers left.

The results from the association analysis provide interesting perspectives. The lift and lift ratio indicate the occurrence of only 14.44% of all transactions where someone under 26 years old buys a commodity outside the highest price segment – this is higher than average. This is explained by the presence of the association rules, which are only relevant for lower-priced premiums, with a lift of 1004.4 for the rules that only involve persons under 26 years. Body repair can be part of the commodity for the under-26

segment, and this is indeed shown by the association rule with a lift of 1.377. The lift of 5.4 is also very telling from the association rule pointing to a car with an LPG or electric engine connected to a comprehensive policy. Individuals interested in a green lifestyle are thus more likely to pay a premium that would cover incidental damages like vandalism. Just like in the previous segment, the association rule 'main driver father' appears here. Practitioners can go through this extensive list of association rules for a segmented target audience and discern which types of coverage or customer characteristics should be combined into one target audience.

5.2. Practical Implications for Insurance Companies

Practical Implications for Insurance Companies - Insurance firms can opt for several solutions when fueled by the findings of the current study. Firstly, insurance firms might derive competitive advantages from offering customers tailor-made services. A real-time or parametric insurance product against drone strikes might be complemented with data on when the swarm of bees was registered in relation to flight paths to adjust premium calculations to the actual client's risk. Secondly, insurance firms might brand tailored micro-insurance products, thereby fostering customer loyalty and engagement. In this strategy, the insurance coverage might be less a product of risk and societal mutual help concepts and more an add-on to the underlying product, such as the insuring of bikes by a bicycle producer. Thirdly, the findings can help insurance firms convert characteristics of existing client segments into emotional/narrative aspects for insurance advertisements. To practically implement a suitable segmentation for an insurance company, the following steps must be taken into account. First, an AI segmentation of clients should be in place. During the process of implementation, the key difficulty, which may occur in strongly regulated countries, could relate to the requirements concerning the cross-use of data not initially collected for insurance purposes. In the European Union, in particular, General Data Protection Regulations could seem harsh.

Then, it is extremely important to enhance the segmentation practice with a feedback loop, which would allow for a continuous improvement of the resolution of the client's needs. This approach would be particularly legitimate in the insurance sector where long-term relationships exist and the customer base can buy a number of products. Hence, it is very important to increase AI competencies to invest in specialists and the

best sourcing of data for AI predictive analytics solutions on the market. The main challenge in implementing AI predictive analytic solutions refers to credibility as such. To introduce such technology, we first must dispel any doubts and show how the analysis is going to draw accurate predictions based on solid, transparent grounds. The situation is more complex for insurers when they need external sources that would provide them with the possibility to enhance their models by introducing data with compelling information. Issues in the insurers' acceptance and practice are widely known and result from the issues of trust, perception, and the fear of the unknown. The issues appear in respect to AI predictive analytics in the insurance industry on the side of two main stakeholders — customers and regulators.

6. Future Direction

This study attempted to examine the status of AI-driven customer segmentation and discussed potential improvements in the context of the insurance industry. However, there are still many areas that have not been addressed. From a technical point of view, there are various emerging technologies that may result in new ways to fulfill our customer segmentation purpose. They include: (a) various machine learning techniques, which have recently become popular in various fields but have not yet been widely adopted in customer segmentation; (b) real-time data segmentation technologies, which can process identity-level data in real time and provide audience- and segment-level insights.

Additionally, we need to look at future customer segmentation from an integrated, unified steering perspective, which includes advanced analytics within a strategy. From a methodological point of view, new and interdisciplinary factors will contribute to the improvement of segmentation solutions. Specifically, data privacy, data ethics, and law issues in customer segmentation should be considered from an interdisciplinary perspective in order to establish a more effective form of ethical analysis. Such interdisciplinary collaboration will become more and more important, especially when AI technologies are used. This may lead to biased patterns and unequal treatment being employed in segmentation, which is a very serious ethical issue. Finally, it is worth noting that AI-driven customer segmentation solutions should be updated or adapted to any change in target markets, and that the model is not always the best, as it may overfit with data and result in suboptimal results or over-segmentation.

7. Conclusion

This paper has outlined the status quo of AI-driven customer segmentation research and then addresses this issue by using supervised learning applied on empirical data from an insurance company. This research has delved into illustrative examples and the practical implications thereof while also illustrating conclusions in a quantitative manner. The study also serves as a helpful tool in an academic context for students looking to illustrate the methodology used in AI segmentation. Methodologies, algorithms and tools: Regarding the methodology, the research is based on unsupervised learning where the steps regarding data understanding, data cleaning, data reduction and data analysis were provided. In this paper, the research aimed to illustrate the steps applied to compile a more illustrative paper regarding the use of AI-driven customer segmentation to a lay audience. Implications: The study was applied in an illustrative example, although the cleaning part was performed in another environment. The insurance dataset was based on comprehensive and car insurance data for two years. While addressing its limitations, this study and paper delves into the customer segmentation in car insurance in-depth—including its implications, the algorithms used, validation and practical applications. In conclusion, it is increasingly becoming important to improve customer experience through different approaches, one of which includes customer segmentation. The adoption of AI-based solutions can enhance customer segmentation and also allow for the analysis of many other variables in insurance. Thus, AI systems can revolutionize how insurance practices are done right now, and we suggest a deeper study of the potential implications of this work to address privacy and regulations issues—under a different model of release, that uses more anonymization.