

Performing Psychographic Segmentation Based on Customer Behaviour

Priyatham Sai Chand Bazaru

Computer Science Department

California State Polytechnic University

Pomona

pbazaru@cpp.edu

Ritika

Computer Science Department

California State Polytechnic University

Pomona

nritika@cpp.edu

Siva Charan Mallena

Computer Science Department

California State Polytechnic University

Pomona

smallena@cpp.edu

Abstract—Customer personality is assessed to find the ideal customer for a business. This is analyzed by using the dataset that contains various attributes pertaining to an individual and what makes him/her an ideal customer. This paper offers a solution how to find such customers by analyzing various aspects with respect to an individual. Customers are clustered into various groups based on attributes and recognized how they can be targeted. Various metrics and types of clustering were used in order to identify such clusters and how the customers in the data can be fit accordingly. It helps a business to modify its product based on its target customers from different types of customer segments. The results obtained from the clustering has shown 3 significant clusters in which the customers can be clustered and how the features have effected the clusters.

Index Terms—attributes, ideal, clustered, metrics, segments

I. INTRODUCTION

In order to stay competitive in the open market, businesses today require them to understand customers and their behaviors thoroughly so they can produce even better products and tailor experiences. This requires analysis of customer's personality and concerns of different types of customers. Using Data Science, Artificial Intelligence and Machine Learning, Individuals can be classified based on their personality traits. Every human being is unique in their own right. Due to multi-modal data capture of users' online activities, a huge corpus of data enabled increased effectiveness of marketing campaigns on a per individual basis like click through rate, resulting in more revenue. The personalized marketing messages and communications are shown to be highly effective and increase not only the demand of products, but also their usage and customer satisfaction. Since recommender systems also depend heavily on knowing each customer's personality, this analysis can be of use in such systems like product recommendations or music recommendations. The predictive relationship of a large and comprehensive set of personal descriptors to aspects of product and brand use is examined. The descriptors comprise demographic and general psychographic variables frequently used in segmentation studies and studies of consumer purchase behavior. The evidence is overwhelming that the covariates are related to brand use in an identical way for all brands, indicating that they are not useful for predicting relative brand preference. The covariates are shown to be predictive

of product use. Discussion of the explanatory content of the variables is offered. [1]

II. DATASET TO BE USED

The dataset consists of 2240 instances upon which clustering is performed. Based on the given data, Clustering of clients in dataset, we will define the segments of the clients by using 4 equally weighted customer segments such as:

- **Stars:** Old customers with high income and high spending nature
- **Need Attention:** New customers with below-average income and low spending nature.
- **High potential:** new customers with high income and high spending nature.
- **Leaky bucket:** old customers with below-average income and a low spending nature.

The dataset defined the following attributes to identify and segment the customers:

- ID: Customer's unique identifier
- YearBirth: Customer's birth year
- Education: Customer's education level
- MaritalStatus: Customer's marital status
- Income: Customer's yearly household income
- Kidhome: Number of children in customer's household
- Teenhome: Number of teenagers in customer's household
- DtCustomer: Date of customer's enrollment with the company
- Recency: Number of days since customer's last purchase
- Complain: 1 if the customer complained in the last 2 years, 0 otherwise

Products

- MntWines: Amount spent on wine in last 2 years
- MntFruits: Amount spent on fruits in last 2 years
- MntMeatProducts: Amount spent on meat in last 2 years
- MntFishProducts: Amount spent on fish in last 2 years
- MntSweetProducts: Amount spent on sweets in last 2 years
- MntGoldProds: Amount spent on gold in last 2 years

Promotion

- NumDealsPurchases: Number of purchases made with a discount

- AcceptedCmp1: 1 if customer accepted the offer in the 1st campaign, 0 otherwise
- AcceptedCmp2: 1 if customer accepted the offer in the 2nd campaign, 0 otherwise
- AcceptedCmp3: 1 if customer accepted the offer in the 3rd campaign, 0 otherwise
- AcceptedCmp4: 1 if customer accepted the offer in the 4th campaign, 0 otherwise
- AcceptedCmp5: 1 if customer accepted the offer in the 5th campaign, 0 otherwise
- Response: 1 if customer accepted the offer in the last campaign, 0 otherwise

Place

- NumWebPurchases: Number of purchases made through the company's website
- NumCatalogPurchases: Number of purchases made using a catalogue
- NumStorePurchases: Number of purchases made directly in stores
- NumWebVisitsMonth: Number of visits to company's website in the last month

Target

Need to perform clustering to summarize customer segments.

III. METHODOLOGY

In this process we separated customers based on common characteristics such as demographics, behaviors or purchasing habits to get a better understanding of the market more effectively. We went over the data in detail like which attributes are more important or discard any attributes not necessary for the project. For example feature like marital status was discarded because there was linear correlation between single and partner income.

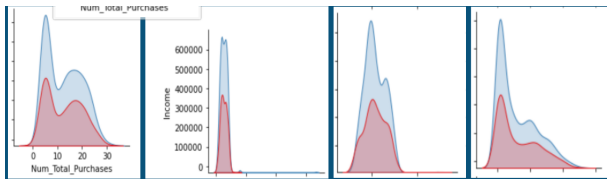


Fig. 1. Graph shows linear correlation between single and married

We also checked for null/not null values instances. Outliers was handled carefully for features like age and income and also data was normalized to make in standard form. In clustering models these unsupervised learning methods we drew references from datasets consisting of input data without labeled responses. To look at the clustering of clients in the dataset, we defined the segments of the clients. Based on spending of customers, segments were created like non-buyer, low-buyer, frequent-buyer, biggest-buyer.

Various Clustering algorithms were used in order to create multiple clusters based on the different attributes that are present to better segment the customers. Algorithms such as the K-means as it is simple and very fast, so in many practical

applications, the method proved to be a very effective way that produced good clustering results. It was also suitable for producing globular clusters [2], Agglomerative clustering since this is a pair-group method: at each iteration exactly two clusters were agglomerated into a single cluster [3]. This was very helpful in our case with multiple attributes where many clusters can be formed. To get a better understanding of the data and to produce a more generalized view with fewer clusters was possible with this method, spectral clustering because Spectral Clustering solely related to the number of data points, but has nothing to do with the dimension [4] was potentially used after pre-processing has been performed on the given data. It is noteworthy that all these algorithms requires number of clusters be defined beforehand. So it was safe to assume that the number of clusters would be above 2 but not more than 8 to be computationally capable to train the data. In our case we had three clusters.

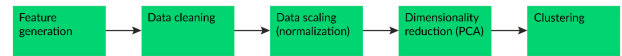


Fig. 2. Methodology flow chart diagram

IV. RESULTS

Multiple algorithms such as the K-means++, Agglomerative clustering and Spectral clustering has resulted in various results on what should be the best number of clusters to differentiate various customers. The following paper has run various tests on metrics such as the silhouette score and sum of squared error (SSE) to find the appropriate number of clusters.

According to our analysis, K-means has not shown any significant peak when measuring the silhouette score and has shown a downward graph. The results of the spectral clustering has shown a similar trend as well. But the SSE with respect to the number of clusters has shown the shape of an elbow when the total number of clusters is 3 providing evidence to our conclusion made by the agglomerative approach.

- Among the 3 models, we saw good metric results for Agglomerative and Spectral Clustering algorithms
- Best Model is AgglomerativeClustering (n = 3) substantiated by the metric of silhouette score with the significant peak formed.

The graph above shows the clear distinction in choosing the 3 clusters to be appropriate and analysing the three clusters with respect to how all the features are distinguished can be seen here. [7] This has shown the final result on what basis are the clusters modeled.

Following are the 3 clusters resulted from Agglomerative Clustering:

Cluster 1:

- Average income = 50000
- Average age = 52 years
- Education = (Graduation, 2n Cycle, Master, PhD)

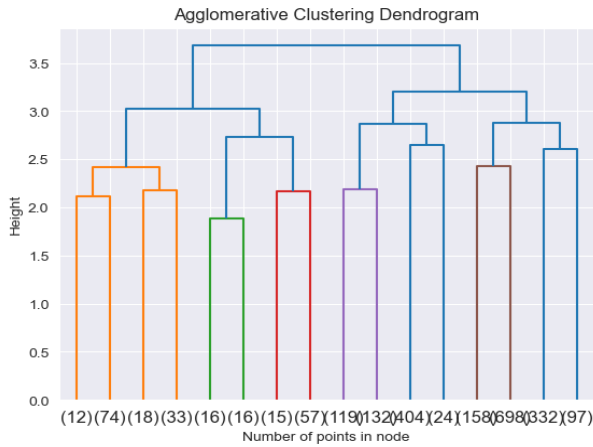


Fig. 3. The dendrogram for agglomerative clustering with no distance threshold. The dendrogram shows the various hierarchical clusters that have been formed along the entire length of the tree. The *distance_threshold* has been applied to obtain the entire tree irrespective of the minimum and maximum clusters defined previously. This shows the various nodes could be used to form two main clusters which is not accurate considering a strong cluster is formed at a height 2.5 which lead us into looking deeper into the number of clusters by using the silhouette score.

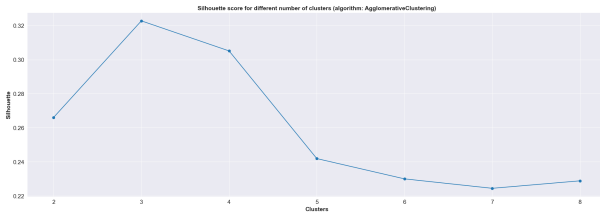


Fig. 4. Silhouette score for different number of clusters by agglomerative method

- People with/without family, families with/without children
- Quite often buy wines also often buy meat
- Most often make purchases on the web
- Average number of purchases = 13

Cluster 2:

- Average income = 70000
- Average age = 55 years
- Education (Graduation, 2n Cycle, Master, PhD)
- People with family with children (Teenhome)
- Quite often buy wines, but they also often buy meat

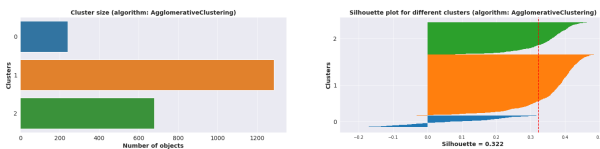


Fig. 5. Agglomerative clustering with $n = 3$

- Most often make purchases in the stores themselves
- Most often make purchases (compared to other clusters)

Cluster 3:

- Average income = 38000
- Average age = 49 years
- Have an education (Graduation, 2n Cycle, Master, PhD)
- People with families with and without children
- Low number of purchases and, accordingly, spend little money on purchases

V. RELATED WORKS

Unlike traditional statistical methods, which rely on domain experts to create hand-crafted features, this paper proposes an unsupervised end-to-end model that can directly and accurately process questionnaire data without human intervention. This paper also demonstrates how to practically apply an automatic unsupervised analysis method (PCA) to find inferences in the big data, and helps interpret the implied meaning to the questions. [5]

Paper offers a primer on machine learning for behavioral scientists. It provides a high-level overview of the type of data typically used in different ML methods, followed by more detailed descriptions of well-established ML methodologies (e.g., supervised learning, unsupervised learning, semi-supervised learning) that can be of value to behavioral researchers. After establishing this technical background, this offer detailed examples of how various ML methodologies can be used for specific behavioral research topics. [6]

VI. CONCLUSION

While it is known for a long time that targeting marketing messages tailored for specific audiences yielded better results, the job of grouping customer base into cohorts has always been done manually by professionals which was time consuming and expensive. This process is naturally not scalable to a company which grew in size and whose goal is of increasing revenue at a rapid rate required adapting to newer customer base. The manual approach is also a bottle neck when introducing new products into the market and designing a marketing campaign for them.

To address the concerns of scalability and also to reduce expenses in terms of money and time, a company's own customer data can be used to generate customer segments for which marketing messaging can be designed for. Clustering algorithms like K-Means, Agglomerative, Spectral clustering can be put to good use to get insights into existing data. We saw that for this dataset, agglomerative clustering with $n = 3$ yielded best results, k-means could also yield good results for a different data set and hence should not be ignored.

It is recommended to run this analysis on customer data regularly on fresh, incoming customer data with clustering algorithms to keep business intelligence updated with any potential changes to customer purchasing behavior with evolving free market.

REFERENCES

- [1] Fennell, G., Allenby, G.M., Yang, S. et al. The Effectiveness of Demographic and Psychographic Variables for Explaining Brand and Product Category Use. *Quantitative Marketing and Economics* 1, 223–244 (2003). <https://doi.org/10.1023/A:1024686630821>
- [2] S. Na, L. Xumin and G. Yong, "Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm," 2010 Third International Symposium on Intelligent Information Technology and Security Informatics, 2010, pp. 63-67, doi: 10.1109/IITSI.2010.74.
- [3] Day, W.H.E., Edelsbrunner, H. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of Classification* 1, 7–24 (1984). <https://doi.org/10.1007/BF01890115>
- [4] Shifei Ding, L. Zhang and Y. Zhang, "Research on Spectral Clustering algorithms and prospects," 2010 2nd International Conference on Computer Engineering and Technology, 2010, pp. V6-149-V6-153, doi: 10.1109/ICCET.2010.5486345.
- [5] Kuo Chi-Hsien and Shinya Nagasawa (2019) Applying machine learning to market analysis: Knowing your luxury consumer, *Journal of Management Analytics*, 6:4, 404-419, DOI: 10.1080/23270012.2019.1692254
- [6] Hagen, L., Uetake, K., Yang, N. et al. How can machine learning aid behavioral marketing research?. *Mark Lett* 31, 361–370 (2020). <https://doi.org/10.1007/s11002-020-09535-7>
- [7] Bazaru et al. "customer-data-analysis" [Source Code] (Version 1.0) (2022) <https://github.com/Priyatham-sai-chand/customer-data-analysis>

VII. SUPPLEMENTARY MATERIAL

Memory Optimization: We performed analysis on mapping the features with datatypes to maintain memory usage.

- Memory usage of Dataframe is 0.50 MB
- Memory usage after optimization is: 0.11 MB
- Decreased by 78.0

Technique taken from : <https://www.kaggle.com/code/alisultanov/clustering-customer-personality-analysis>

The link to the Dataset (marketing_campaign.csv): <https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis> The link to the repository of code that is used: <https://github.com/Priyatham-sai-chand/customer-data-analysis>