JOURNAL OF MODERN SCIENCE
SPECIAL ISSUE
3/57/2024
www.jomswsge.com

DOI: doi.org/10.13166/jms/191180

SYLWESTER BOGACKI
WSEI University in Lublin, Poland
ORCID iD: orcid.org/0000-0002-8330-4573

PAWEŁ RYMARCZYK
Netrix S.A., Poland
ORCID iD: orcid.org/0000-0002-5990-4771

TOMASZ SMUTEK
WSEI University in Lublin, Poland
ORCID iD: orcid.org/0009-0003-4756-7354

MAREK RUTKOWSKI
WSEI University in Lublin, Poland
ORCID iD: orcid.org/0009-0009-5910-7893

AGNIESZKA CHMIELOWSKA-MARMUCKA
Graduate School of Business - National-Louis University, Poland
ORCID iD: orcid.org/0000-0003-0105-4038

ADVANCED METHODS FOR TARGET AUDIENCE IDENTIFICATION: ENHANCING MARKETING STRATEGIES THROUGH MACHINE LEARNING AND DATA ANALYTICS

ZAAWANSOWANE METODY IDENTYFIKACJI ODBIORCÓW DOCELOWYCH: ULEPSZANIE STRATEGII MARKETINGOWYCH POPRZEZ UCZENIE MASZYNOWE I ANALIZĘ DANYCH

Abstract

Purpose: This article presents a novel approach that leverages advanced data analytics and machine learning techniques to enhance marketing strategies. By precisely targeting and segmenting audience groups based on their descriptive profiles, the study aims to significantly improve the efficacy of marketing campaigns.

Methods: The study employs several clustering and community detection algorithms, including Louvain Community, Greedy Modularity, and Label Propagation. These methods are applied to diverse datasets to identify distinct groups within the audience that exhibit specific behavioral and preference patterns. The approach emphasizes data-driven decision-making, which involves making decisions based on the analysis of data rather than intuition or observation, to optimize marketing outcomes.

Results: demonstrate that employing advanced clustering techniques can significantly refine the segmentation process, leading to more targeted marketing efforts. These methods successfully identified nuanced sub-groups within the datasets, which corresponded closely with customer behaviors and preferences variations, thereby allowing for more tailored marketing strategies.

Discussion: The study's findings underscore the imperative for marketers to embrace sophisticated analytical techniques. Machine learning has the potential to transform marketing strategies by providing deeper insights into customer segmentation. This research highlights the importance of staying ahead of the curve in the face of the complexities of consumer markets and evolving business environments.

Streszczenie

Cel: Artykuł bada zastosowanie zaawansowanych technik analityki danych i metod uczenia maszynowego w celu poprawy strategii marketingowych poprzez dokładne targetowanie i segmentację grup odbiorców na podstawie ich opisów profilowych.

Metody: Badanie wykorzystuje kilka algorytmów klasteryzacji i wykrywania wspólnot, w tym Louvain Community, Greedy Modularity i Label Propagation. Metody te stosowane są do różnorodnych zbiorów danych, aby zidentyfikować wyraźne grupy wśród odbiorców, które wykazują specyficzne wzorce zachowań i preferencji. Podejście to kładzie nacisk na podejmowanie decyzji opartych na danych w celu optymalizacji wyników marketingowych.

Wyniki: Wyniki pokazują, że zastosowanie zaawansowanych technik klasteryzacji może znacząco usprawnić proces segmentacji, prowadząc do bardziej ukierunkowanych wysiłków marketingowych. Metody te z powodzeniem zidentyfikowały subtelne sub-grupy w zbiorach danych, które ściśle odpowiadają różnicom w zachowaniach i preferencjach klientów, co pozwala na bardziej spersonalizowane strategie marketingowe.

Omówienie: Wyniki podkreślają potencjał uczenia maszynowego w transformacji strategii marketingowych, dostarczając głębszych wglądów w segmentację klientów. Badanie podkreśla znaczenie przyjęcia zaawansowanych technik analitycznych, aby nadążyć za złożonościami rynków konsumenckich i ewoluującymi środowiskami biznesowymi.

Keywords: *clustering, machine learning, data analysis*

Słowa kluczowe: *segmentacja, uczenie maszynowe, analiza danych*

Introduction

In the era of advancing data analytics, a precise understanding and segmentation of the target audience has become a pivotal component of marketing strategies. Target audience identification and segmentation enhance the precision of marketing campaigns and enable marketers to understand consumer behavior and needs more profoundly. Enterprises strive to tailor their products and services more closely to customer needs. In this context, methods for identifying target groups based on their descriptions play a central role (Galiano Coronil, 2022; Omidvar-Tehrani et al., 2019).

Integrating artificial intelligence within marketing frameworks is crucial for achieving more nuanced audience segmentation and precise target identification. This approach is essential for customizing marketing efforts and enhancing customer engagement. The transformative role of advanced data analytics in decoding complex consumer data into actionable insights is significantly highlighted in contemporary research. This integration facilitates a deeper understanding of customer behavior, enabling marketers to tailor their strategies effectively (Haleem et al., 2022; Huang & Rust, 2021; Mandapuram et al., 2020). Furthermore, existing literature provides a comprehensive framework demonstrating how big data analytics facilitates the finer segmentation of customer bases and the identification of key market segments. These data-driven insights are integral to developing strategic marketing practices catering to consumer needs and preferences. The ability to analyze and apply vast amounts of data reflects a significant advancement in marketing strategies, aligning theoretical foundations with practical applications to optimize marketing outcomes (Grover et al., 2018; Tam et al., 2021; Yoseph et al., 2020). Furthermore, the development of integrated machine learning systems for analyzing multifaceted data sources, as explored, significantly enhances the capability of marketers to adapt and innovate in creating business processes that cater precisely to customer demands (Rymarczyk,

Bednarczuk, et al., 2021). In addition, the application of modern machine learning techniques for customer profiling and segmentation, as investigated, underlines the importance of employing advanced computational methods, such as the GRU network, to effectively process and analyze customer data (Rymarczyk, Golabek, et al., 2021).

In recent years, technological advances have greatly improved the ability to handle and analyze large amounts of data. Tools such as AutoEmbedder and Principal Component Analysis (PCA) have emerged as valuable resources for transforming categorical variables into vector spaces, allowing more effective data analysis techniques. AutoEmbedder, for example, is a recent innovation that adapts embedding methods to handle categorical data more efficiently. This approach not only preserves the inherent relationships within the data, but also minimizes issues related to high dimensionality, a common challenge in data analysis (Rachwał et al., 2023). By embedding categorical variables, researchers can reduce the number of dimensions while retaining the essential information, making the data more manageable and the analysis more accurate. Conversely, Principal Component Analysis (PCA) is a widely recognized statistical method designed to highlight variations and expose distinct patterns within a dataset. This technique transforms the original set of variables into new variables, linear combinations of the originals. Known as principal components, these new variables are selected to maximize variance, thereby offering a strategy to reduce data complexity with minimal information loss (Abdulhafedh, 2021).

The advent of machine learning in customer data analysis has significantly advanced the capabilities of businesses to understand and cater to their diverse customer base. Machine learning algorithms, particularly those focused on clustering, have become instrumental in discovering patterns and groupings within large datasets that traditional methods could not easily discern. Several clustering algorithms have been spotlighted in literature for their efficacy in customer segmentation. Algorithms like K-means and DBSCAN are praised for their simplicity and effectiveness in handling vast datasets. For instance, Hicham and Karim highlight the use of clustering ensemble techniques for more efficient customer segmentation, suggesting a combination of DBSCAN, K-means, MiniBatch K-means, and mean-shift algorithms for optimal results

(Hicham & Karim, 2022). Similarly, Hung et al. discuss the application of hierarchical agglomerative clustering (HAC) in segmenting customer data, demonstrating its potential to reveal meaningful customer groups based on purchasing behavior and preferences (Hung et al., 2019).

Furthermore, newer methodologies such as the Louvain Community Detection and Greedy Modularity techniques are also being explored (Gonzalez-Montesino et al., 2023; Rustamaji et al., 2024). These methods are noted for their ability to detect community structures within networks, which can be analogous to identifying customer segments with similar traits or behaviors. This innovative clustering approach can uncover subtler and more complex patterns that traditional methods might miss (Zatonatska, Liashenko, Feraniuk, Skowron, Wołowiec, Dluhopolskyi, 2023).

The evaluation of clustering methods using various indices and metrics has emerged as a significant area of research, particularly given the diversity and complexity of datasets in fields like bioinformatics, social network analysis, and machine learning. Notable among these metrics are the Caliński-Harabasz index, the Davies-Bouldin index, and modularity measures, which offer insights into the quality of clustering outcomes. This discussion is based on comparing different clustering algorithms, emphasizing how these metrics reflect the effectiveness of each method across various datasets. The Caliński-Harabasz index is a well-regarded measure that evaluates cluster validity by comparing the sum of between-cluster dispersion to within-cluster dispersion, favoring models that are tightly grouped internally and well-separated externally. Conversely, the Davies-Bouldin index measures the average similarity between clusters, where lower values indicate more distinctly separated clusters. Modularity, on the other hand, is crucial in network analysis and assesses the strength of the division of a network into modules, thus evaluating the non-random structure of network clusters(Bihari et al., 2024; Daraghmeh et al., 2023; Wei et al., 2021).

Research has demonstrated varied responses to these metrics using different clustering techniques. For example, K-means and hierarchical clustering have been shown to perform differently across these indices when applied to genetic data or social network (Halim et al., 2021; Lu et al., 2023)This variability underscores the necessity of selecting appropriate clustering methods based on the dataset characteristics and the specific goals of the analysis.

Combining these advanced tools allows researchers to conduct more nuanced analyses of target groups, leading to better decision-making and more tailored strategy development. Integrating such methodologies into the research process highlights the evolving nature of data analysis, which is increasingly moving towards automation and high-dimensional data handling. Influential audience targeting based on detailed target descriptions requires a combination of methodological approaches capable of providing deep insights into the complexity of customer data. The results of such approaches are invaluable because they enable more targeted and efficient marketing strategies, thereby increasing business effectiveness (Zhuravka, Filatova, Petr Šuleř, Wołowiec 2021). The article aimed to examine the effectiveness of advanced data analytics techniques and machine learning methods in enhancing marketing strategies by identifying and segmenting target groups based on their descriptions. The study focused on utilizing various clustering and community detection methods, such as Louvain Community, Greedy Modularity, Label Propagation, K-means, and DBSCAN, to segment datasets into meaningful groups that reflect the nuances of customer behaviors and preferences.

Research Methodology

The research was conducted using advanced data analysis techniques and machine learning algorithms. The process began with coding customer data using the AutoEmbedder tool, which allowed categorical variables to be transformed into a vector space. This permitted further data analysis using the Principal Component Analysis (PCA) method, which preserved 95% of the variance in the data. The next step was to generate a similarity matrix based on data embeddings. On this basis, a graph of customer relationships was built using selected cutoff parameters. These graphs allowed the implementation of clustering algorithms based on graph theory, such as Louvain Community, Greedy Modularity, and Label Propagation methods.

Cosine similarity ranges from 0 to 1 and was selected as the metric for assessing similarity. Various techniques were employed to categorize customers into groups based on similar characteristics, including K-means and DBSCAN, along with graph-based methods such as Louvain Community, Greedy Modularity,

and Label Propagation. These methods necessitated the construction of a graph that depicted the interactions among the most similar customers within the dataset. This graph was required to be connected. The graph's structure was derived from a similarity matrix, establishing links between customers regulated by a threshold value. Links were created between customers whose similarity scores met or exceeded predetermined threshold levels set at 0.25, 0.5, and 0.75.

The Louvain Community method, a well-known community detection algorithm, optimizes modularity over multiple stages, progressively merging nodes to form larger communities. The code provided applies this method to a graph constructed from a cosine similarity matrix, coloring nodes in a visualization based on their community membership. This is achieved using NetworkX and a color map scaled to the identified distinct communities, allowing for a visual representation of the community structure.

Similarly, the Greedy Modularity method is used for community detection, focusing on maximizing the modularity score directly and greedily. Starting with each node as a separate community, it iteratively connects pairs of communities that provide the highest increase in modularity until no further improvements can be made. Nodes in the resulting graph are visualized with colors corresponding to their community assignments as determined by the method.

The Label Propagation method operates dynamically, where labels representing community identifiers are propagated through the network. Each node adopts the most common label among its neighbors, leading to rapid local consensus and eventual convergence to a coherent community structure. The graph is visualized after community detection, with nodes colored by their community labels reflecting the clusters identified by this process.

DBSCAN groups points that are densely packed together while marking points in low-density regions as outliers. This clustering method is parameter dependent, requiring an eps (the maximum separation distance between two tests so that one can be taken in the vicinity of the other) and min_samples (the number of specimens in the area for a given point to be recognized as a critical feature). The provided code explores multiple settings for these parameters to find the best clustering configuration, which is evaluated using silhouette scores. The results of DBSCAN are further visualized by calculating the number of clusters and noise points, which helps to identify density-based clustering.

The K-means clustering algorithm was applied to divide the data into k clusters, with each data point assigned to the cluster whose centroid is closest. Thus, centroids are located, and points are grouped accordingly. This iterative refinement of centroids continues until their positions no longer change significantly. A visualizer tool is used to determine the best number of clusters by employing the silhouette score, which helps pinpoint the appropriate k-value for the dataset.

Cluster results were evaluated using various unsupervised clustering metrics. The Calinski-Harabasz and Davies-Bouldin indices were utilized to determine the most effective clustering solution. The Normalized Mutual Information (NMI) measure and the Fowlkes-Mallows index were applied to assess the similarity between two sets of clusters. Modularity levels were also compared for methods based on graphs. The optimal cluster number for the K-means algorithm was determined through the elbow method, silhouette scores, and the values obtained from the Calinski-Harabasz and Davies-Bouldin indicators. Parameters for the DBSCAN clusters were chosen based on the silhouette results, which guided the clustering process effectively.

In the framework of the conducted studies, two data sets were analyzed. The first set, containing information on 701 commercial intermediaries, comprised columns providing data about the identifier of the intermediary, business type, number of employees, frequency of orders, timing of orders, and financial and logistical data. Additionally, information on the city's population and the region's GDP per capita was appended to this set based on variables related to the city and state or province names. All continuous variables in the dataset were categorized based on histograms, intended to facilitate their further processing using the AutoEmbedder tool.

The second dataset consisted of information about 2240 retail customers, including their identifiers, year of birth, marital status, education level, number of children, annual household income, and purchasing behaviors such as complaint frequency, expenditures on various product categories, and the manner and frequency of purchases. The frequency of customer website visits was also taken into account. It was noted that the dominant group within the dataset were customers in marital relationships and with higher education, which could influence the grouping results.

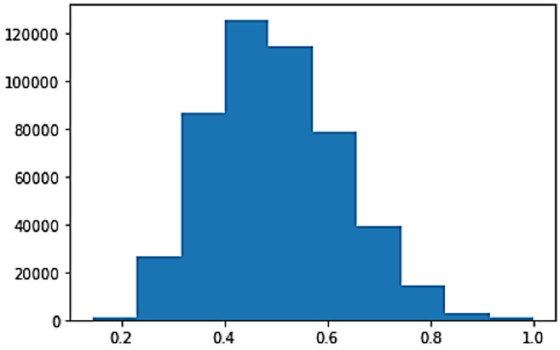
Clustering on a set of intermediaries

The clustering algorithms were tested on a dataset of brokers using the AutoEmbedder. The AutoEmbedder did not generate errors during the entire dataset encoding, as shown in Figure 1. After applying the AutoEmbedder, the dataset consisted of 132 columns, subsequently reduced to 43 using PCA. Similarity between clients was determined based on cosine similarity, and the distribution of scaled similarities is shown in Figure 2.

Figure 1. *Percentage of AutoEmbedder errors when coding the test set*

	Feature	Percent wrong encoding
0	BusinessType	0.0%
1	OrderFrequency	0.0%
2	OrderMonth	0.0%
3	FirstOrderYear	0.0%
4	LastOrderYear	0.0%
5	ProductLine	0.0%
6	AnnualSales	0.0%
7	MinPaymentType	0.0%
8	MinPaymentAmount	0.0%
9	AnnualRevenue	0.0%
10	YearOpened	0.0%
11	EnglishCountryRegionName	0.0%
12	GDPBins	0.0%
13	PopulationBins	0.0%
14	AvgTicketsBikesBins	0.0%
15	AvgTicketsClothingBins	0.0%
16	AvgTicketsAccessoriesBins	0.0%
17	NumberEmployeesBins	0.0%
18	AvgTicketsSeasonalDiscountBins	0.0%
19	AvgTicketsExcessInventoryBins	0.0%
20	AvgTicketsDiscontinuedProductBins	0.0%
21	AvgTicketsVolumeDiscountBins	0.0%
22	AvgTicketsNewProductBins	0.0%
23	AvgTicketsNoDiscountBins	0.0%
24	AvgDaysNewProductBins	0.0%
25	StdDaysNewProductBins	0.0%
26	MinDaysNewProductBins	0.0%
27	MaxDaysNewProductBins	0.0%
28	AvgFreightBins	0.0%
29	StdFreightBins	0.0%
30	MinFreightBins	0.0%
31	MaxFreightBins	0.0%

Figure 2. *Distribution of rescaled cosine similarities over a set of intermediates*



The initial threshold was set at 0.75, which resulted in a highly fragmented graph with over 400 potential clusters for only 701 observations, making it unsuitable for clustering. Reducing the cutoff to 0.5 enabled the Louvain Community method to identify three distinct groups: one with 151 intermediaries, another with 215, and the largest with 335 intermediaries. Meanwhile, the Greedy Modularity approach divided the data into two almost evenly sized groups of 351 and 350 intermediaries, respectively, and the Label Propagation method also identified two groups, each very similar in size.

When the Fowlkes-Mallows index was used to compare the clustering results, it suggested that the clusters formed by the Greedy Modularity and Label Propagation methods were identical, and those by the Louvain Community method were pretty similar, with a similarity coefficient of 0.83. The Normalized Mutual Information (NMI) values confirmed these findings, indicating exact similarity between the Greedy Modularity and Label Propagation clusters and a slightly less, yet still significant, similarity for the Louvain Community method clusters.

Further investigation with a lower cutoff of 0.25 showed the Louvain Community method dividing the clients into two groups, one with 356 and another with 345 intermediaries. Greedy Modularity formed two groups with 382 and 319 intermediaries, respectively. The Label Propagation method, however, identified just one group comprising all intermediaries. At this threshold, the groups identified by the Label Propagation method were moderately similar to those from the other

methods, according to the Fowlkes-Mallows index. Still, they were the least similar when comparing the Louvain Community and Greedy Modularity methods. The NMI scores indicated that the clusterings were entirely dissimilar. Modularity values were notably low, indicating a weak community structure.

In all cases, the DBSCAN method resulted in a single cluster, leaving some clients unclassified. The analysis of the clustering methods using the Calinski-Harabasz and Davies-Bouldin indices showed better scores for clusterings with the AutoEmbedder, with the former indicating higher cluster separation and the latter showing lower within-cluster dispersion.

Table 1. *Values of Calinsky-Harabash and Davies Bouldin indices for divisions of a set of intermediaries*

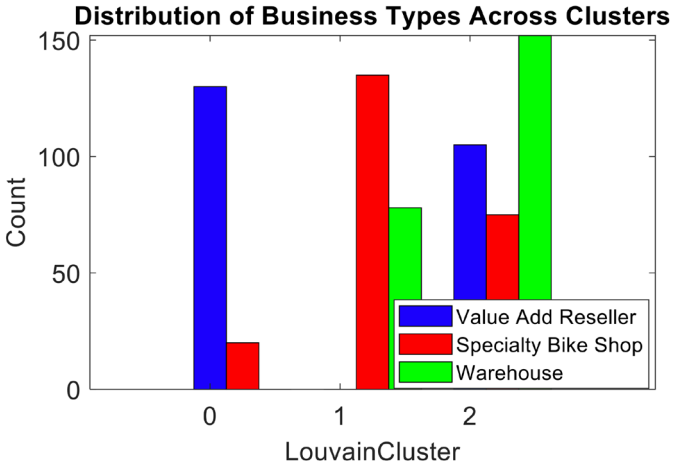
Parameter cut-off	Method	Calinsky Harabash Index	Davies Bouldin Index
0.5	Louvain Community	44.240	3.784
0.5	Greedy Modularity	69.146	3.163
0.5	Label Propagation	69.146	3.163
-	DBSCAN	59.238	2.869
0.25	Louvain Community	68.947	3.166
0.25	Greedy Modularity	22.982	5.445
0.25	Label Propagation	-	-

Table 2 displays the modularity values for various graph-based methods at different cutoff levels, excluding scenarios where only a single group was formed. The Greedy Modularity and Label Propagation methods achieved the highest modularity scores, particularly at a cutoff of 0.5 and using AutoEmbedder for clustering. These results indicate that the clusters formed are notably distinct, particularly in the variety of businesses within each group, as illustrated in Figure 3.

Table 2. *Modularity parameters for different graph algorithms and different cutoff parameters on the set of intermediates*

Parameter cut-off	Method	Modularity
0.5	Louvain Community	0.278
0.5	Greedy Modularity	0.290
0.5	Label Propagation	0.290
0.25	Louvain Community	0.086
0.25	Greedy Modularity	0.053

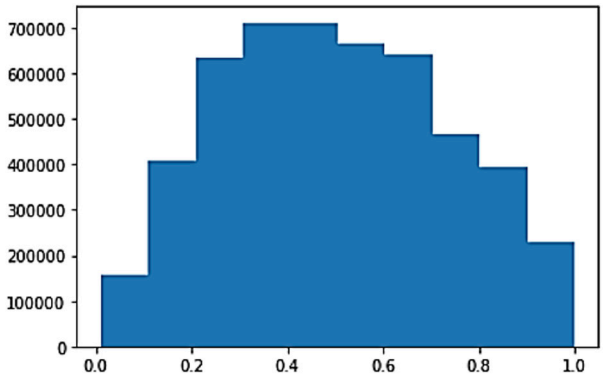
Figure 3. *Types of enterprises in groups by Louvain Community method, cosine similarity*



Clustering results in a set of retail customers

The AutoEmbedder was used to encode a dataset without generating any errors. After applying the AutoEmbedder, the dataset consisted of 96 columns, subsequently reduced to 8 columns using PCA. The similarities among customers were assessed using the cosine similarity metric, and the results of this analysis, after scaling, are presented in Figure 5.

Figure 5. *Distribution of rescaled cosine similarities over a set of retail customers*



Using a cutoff parameter of 0.75, three groups were identified by the Louvain Community method: Group 0 contained 565 clients, Group 1 contained 929 clients, and Group 2 contained 746 clients. The Greedy Modularity and Label Propagation methods split the data set into two groups. In the case of Greedy Modularity and Label Propagation, Group 0 contained 1285 clients, while Group 1 contained 955 clients. The partitioning achieved by these two methods was similar, with the Fowlkes-Mallows index yielding a value of approximately 0.98 and the NMI parameter yielding a value of roughly 0.94. When the cutoff parameter was changed to 0.5, the Louvain Community method also identified three communities, but the distribution across the groups was not as uniform as with a cutoff of 0.75. The Greedy Modularity method produced two groups: the first with 1200 clients and the second with 1040 clients. The Label propagation method found only one community that included all clients. With a cutoff parameter of 0.25, the Louvain Community

method divided the dataset into two groups: Group 0 with 1048 clients and Group 1 with 1192 clients. Similarly, the Greedy Modularity method produced two groups: Group 0 with 1131 clients and Group 1 with 1109 clients. The Label Propagation method consolidated all clients into a single community.

For graph-based clustering methods, comparisons were made using the modularity score. The division produced by the Louvain Community method, particularly at a cutoff of 0.75, displayed the highest modularity, recorded at 0.5656527057449383. The DBSCAN method identified three groups: Group 0 included 950 clients, Group 1 included 831 clients, and Group 2 included 16 clients, leaving 443 unclassified. The values of the Calinski-Harabasz and Davies-Bouldin indices for different methods and cutoff parameters are detailed in Table 3.

Table 3. *Values of Calinsky-Harabash and Davies Bouldin indices for different methods and different cutoff parameters*

Parameter cut-off	Method	Calinsky Harabash Index	Davies Bouldin Index
0.25	Louvain Community	801.783	1.626
0.25	Greedy Modularity	796.954	1.640
0.25	Label Propagation	-	-
0.5	Louvain Community	415.123	2.358
0.5	Greedy Modularity	782.355	1.644
0.5	Label Propagation	-	-
-	Dbscan	314.034	2.484
0.75	Louvain Community	542.745	2.080
0.75	Greedy Modularity	773.752	1.624
0.75	Label Propagation	781.243	1.625

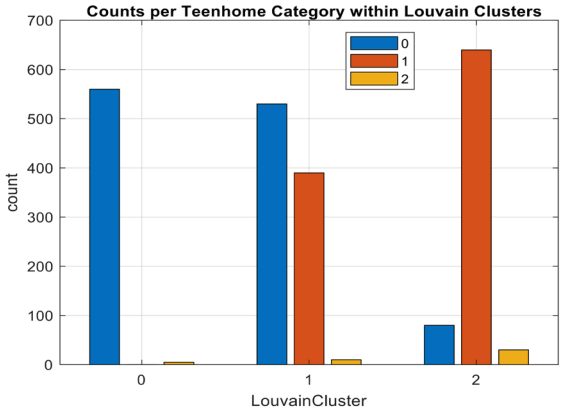
Table 4 presents the Modularity values for different graph-based methods and cutoff parameters. The highest value corresponds to the Louvain Community method, which has a cutoff of 0.75.

Table 4. *Modularity parameters for different graph algorithms and different cutoff parameters on the set of intermediates*

Parameter cut-off	Method	Modularity
0.75	Louvain Community	0.566
0.75	Greedy Modularity	0.463
0.75	Label Propagation	0.409
0.5	Louvain Community	0.309
0.5	Greedy Modularity	0.309
0.25	Louvain Community	0.205
0.25	Greedy Modularity	0.202

The Louvain Community method, with a cutoff parameter of 0.75, resulted in three well-balanced groups. While this division had slightly lower Calinski-Harabasz and Davies-Bouldin index scores compared to those from the Greedy Modularity method at a 0.5 cutoff, it achieved a higher modularity score. This particular segmentation was notably differentiated by the presence of teenagers in the households. Figure 6 illustrates the distribution of teenagers across the groups: households without teenagers in Group 0, those with one teenager in Group 2, and a mix of both in Group 1.

Figure 6 *Groups formed by Louvain's Community method (using AutoEmbedder), breakdown by number of adolescents owned in the household*



Conclusions

The research on identifying target groups based on their descriptions has successfully demonstrated the integration of advanced data analytics and machine learning techniques to improve marketing strategies. This research primarily focused on using various clustering and community detection methods such as Louvain Community, Greedy Modularity, Label Propagation, DBSCAN, and K-means to segment data sets into meaningful groups that reflect the subtleties of customer behaviors and preferences.

Utilizing AutoEmbedder and PCA for data encoding and dimensionality reduction maintained a significant portion of the data's variance, which is crucial for preserving the original data characteristics while simplifying the computational processes.

Applying cosine similarity to developing similarity matrices was pivotal in accurately mapping customer relationships, which is essential for applying graph-based clustering techniques. The research highlighted the importance of parameter selection, such as the cutoff parameter in graph-based methods, which significantly influences the structure and quality of the resulting clusters. This aspect was crucial in optimizing the clustering output to match the predefined user characteristics and expectations better.

Using metrics like Calinski-Harabasz and Davies-Bouldin indices, the research could objectively assess the effectiveness of different clustering methods. These metrics provided a quantitative basis to compare the cohesion and separation of clusters, guiding the selection of the most appropriate clustering method.

The comprehensive analysis and evaluation of different clustering techniques used in the research revealed distinct strengths and suitability of each method depending on the specific data characteristics and research objectives. However, among the various methods tested, the Louvain Community method consistently demonstrated higher effectiveness in forming well-defined, coherent groups, notably aligned with the complex interaction patterns within the datasets.

The Louvain Community method excelled particularly in optimizing modularity, which proved beneficial for detecting communities within large and complex networks. This optimization allowed for a more nuanced dataset segmentation, which is crucial in identifying and understanding subtle customer

behavior patterns. The high modularity scores obtained with the Louvain Community method indicate that it was particularly effective in maximizing intra-cluster similarity while minimizing inter-cluster similarities, leading to more distinct and actionable customer segments.

The effectiveness of this method was confirmed through a comparative analysis using the Calinski-Harabasz and Davies-Bouldin indices. Although the Greedy Modularity and Label Propagation methods also yielded positive results, the Louvain Community method consistently outperformed them, as indicated by higher Calinski-Harabasz indices, which reflect better cluster validity, and lower Davies-Bouldin indices, demonstrating improved separation between clusters.

The insights from the clustering analysis can be directly applied to refine marketing strategies. By understanding the characteristic features of each cluster, businesses can tailor their marketing efforts better to meet each target group's specific needs and preferences. Identifying customer groups with similar behaviors and preferences allows for more personalized customer engagement strategies, enhancing customer satisfaction and loyalty. Insights into different clusters' specific needs and preferences can guide product development and innovation strategies, ensuring that new products align more closely with customer expectations.

References

Abdulhafedh, A. (2021). Incorporating K-means, Hierarchical Clustering and PCA in Customer Segmentation. *Journal of City and Development, Vol. 3, 2021, Pages 12-30*, *3*(1), 12–30. https://doi.org/10.12691/JCD-3-1-3

Bihari, A., Vishwakarma, S., Kumar Bhardwaj, S., Tripathi, S., Agrawal, S., & Joshi, P. (2024). Cancer Gene Clustering Using Computational Model. / *GMSARN International Journal*, *18*, 252–257.

Daraghmeh, M., Agarwal, A., & Jararweh, Y. (2023). An ensemble clustering approach for modeling hidden categorization perspectives for cloud workloads. *Cluster Computing*. https://doi.org/10.1007/S10586-023-04205-5

Galiano Coronil, A. (2022). Behavior as an approach to identifying target groups from a social marketing perspective. *International Review on Public and Nonprofit Marketing*, *19*(2), 265–287. https://doi.org/10.1007/S12208-021-00298-Z/FIGURES/5

Gonzalez-Montesino, L., Grass-Boada, D. H., & Armannazas, R. (2023). Network Community Detection in Connectomics Data using Graph Theory. *Proceedings – 2023 2023 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2023*, 3459–3465. https://doi.org/10.1109/BIBM58861.2023.10385337

Grover, V., Chiang, R. H. L., Liang, T. P., & Zhang, D. (2018). Creating Strategic Business Value from Big Data Analytics: A Research Framework. *Journal of Management Information Systems*, *35*(2), 388–423. https://doi.org/10.1080/07421222.2018.1451951

Haleem, A., Javaid, M., Asim Qadri, M., Pratap Singh, R., & Suman, R. (2022). Artificial intelligence (AI) applications for marketing: A literature-based study. *International Journal of Intelligent Networks*, *3*, 119–132. https://doi.org/10.1016/J.IJIN.2022.08.005

Halim, Z., Sargana, H. M., Aadam, Uzma, & Waqas, M. (2021). Clustering of graphs using pseudo-guided random walk. *Journal of Computational Science*, *51*, 101281. https://doi.org/10.1016/J.JOCS.2020.101281

Hicham, N., & Karim, S. (2022). Analysis of Unsupervised Machine Learning Techniques for an Efficient Customer Segmentation using Clustering Ensemble and Spectral Clustering. *IJACSA) International Journal of Advanced Computer Science and Applications*, *13*(10). www.ijacsa.thesai.org

Huang, M. H., & Rust, R. T. (2021). A strategic framework for artificial intelligence in marketing. *Journal of the Academy of Marketing Science*, *49*(1), 30–50. https://doi.org/10.1007/S11747-020-00749-9/TABLES/4

Hung, P. D., Thuy Lien, N. T., & Ngoc, N. D. (2019). Customer segmentation using hierarchical agglomerative clustering. *ACM International Conference Proceeding Series*, *Part F148384*, 33–37. https://doi.org/10.1145/3322645.3322677

Lu, M., Guo, Z., & Gao, Z. (2023). Effect of intracranial electrical stimulation on dynamic functional connectivity in medically refractory epilepsy. *Frontiers in Human Neuroscience*, *17*, 1295326. https://doi.org/10.3389/FNHUM.2023.1295326/FULL

Mandapuram, M., Mandapuram, M., Gutlapalli, S. S., Reddy, M., & Bodepudi, A. (2020). Application of Artificial Intelligence (AI) Technologies to Accelerate Market Segmentation. *Global Disclosure of Economics and Business*, *9*(2), 141–150. https://doi.org/10.18034/gdeb.v9i2.662

Omidvar-Tehrani, B., Amer-Yahia, S., & Borromeo, R. M. (2019). User group analytics: hypothesis generation and exploratory analysis of user data. *VLDB Journal*, *28*(2), 243–266. https://doi.org/10.1007/S00778-018-0527-4/FIGURES/17

Rachwał, A., Popławska, E., Gorgol, I., Cieplak, T., Pliszczuk, D., Skowron, Ł., & Rymarczyk, T. (2023). Determining the Quality of a Dataset in Clustering Terms. *Applied Sciences 2023, Vol. 13, Page 2942*, *13*(5), 2942. https://doi.org/10.3390/APP13052942

Rustamaji, H. C., Kusuma, W. A., Nurdiati, S., & Batubara, I. (2024). Community detection with Greedy Modularity disassembly strategy. *Scientific Reports*, *14*(1). https://doi.org/10.1038/S41598-024-55190-7

Rymarczyk, P., Bednarczuk, P., Nowak, R., & Cieplak, T. (2021). Methods of Analyzing Consumer Behavior Based on Multi-Source Data. *EUROPEAN RESEARCH STUDIES JOURNAL*, *XXIV*(Special Issue 2), 335–345. https://doi.org/10.35808/ERSJ/2229

Rymarczyk, P., Golabek, P., Sylwia, & Rzemieniak, M. (2021). Profiling and Segmenting Clients with the Use of Machine Learning Algorithms. *EUROPEAN RESEARCH STUDIES JOURNAL*, *XXIV*(Special Issue 2), 513–522. https://doi.org/10.35808/ERSJ/2281

Tam, P. T., Son, D. M., Tan, T. Le, & Ha, H. (2021). Data Driven Customer Segmentation for Vietnamese SMEs in the Big Data Era. *Macro Management & Public Policies*, *3*(2), 33–43. https://doi.org/10.30564/MMPP.V3I2.3553

Wei, J., Ma, H., Liu, Y., Li, Z., & Li, N. (2021). Hierarchical high-order co-clustering algorithm by maximizing modularity. *International Journal of Machine Learning and Cybernetics*, *12*(10), 2887–2898. https://doi.org/10.1007/S13042-021-01375-9/TABLES/6

Yoseph, F., Ahamed Hassain Malim, N. H., Heikkilä, M., Brezulianu, A., Geman, O., & Paskhal Rostam, N. A. (2020). The impact of big data market segmentation using data mining and clustering techniques. *Journal of Intelligent & Fuzzy Systems*, *38*(5), 6159–6173. https://doi.org/10.3233/JIFS-179698.

Zatonatska, T., Liashenko, O., Feraniuk, Y., Skowron, Ł., Wołowiec, T., Dluhopolskyi, O. (2023). Impact of Migration on Forecasting Budget Expenditures on Education, Sustainabilit,.15(21), p.15473; https://doi.org/10.3390/su152115473

Zhuravka, F., Filatova, H., Šuleř, P., Wołowiec, T. (2021). State debt assessment and forecasting: time series analysis, Investment Management and Financial Innovations, 18(1), p. 65-75. doi:10.21511/imfi.18(1).2021.06