A Hybrid Approach to Optimize Market Segmentation: Integration of Unsupervised Clustering and Supervised Ensemble Method

Sakshi Dua[1*], Devender Kumar[2], Sonu Dua[3] and Jyoti Batra[4]

[1,2,4]School of Computer Applications, Lovely Professional University, Punjab, India
[3]School of Management, Lyallpur Khalsa College of Technical Campus, Punjab, India
*Corresponding author e-mail id: [1]sakshi_nancy@yahoo.in, [2]devender.kumar2k7@gmail.com, [3]sonudua3778 @gmail.com [4]jyotijenni19@gmail.com

Abstract. Market segmentation is critical process in comprehending customer diversity and customizing marketing techniques for better business outcomes. Using ensembled method upon real time dataset, this research work is presenting a robust framework for market segmentation. The main aim is to search for meaningful and significant customer clusters and correctly classify them so that companies could be enable or implement their marketing and resource allocation practices. The primary dataset collected is comprised of multi-dimensional features, behavioral, and psychographic attributes. At beginning, encoding categorical features, normalization of numerical values using dimensionality reduction such as Principal Component Analysis (PCA) and t-SNE (t-Distributed Stochastic Neighbor Embedding) for feature simplification process are deployed. Clustering is being performed by using Gaussian Mixture Model (GMM) and Silhouette score was useful to optimize cluster numbers and covariance type. For effective allocation of customers to identified segments, a classification model based on ensembled learning method such as XGBoost over primary dataset is being trained by using post clustering. To assess the performance of model, various metrics such as: classification accuracy, silhouette score and visual cluster validation is calculated. Results of research work demonstrates accurate and precise classification system with improved interpretability of customer behavior and presents significant insights for market segmentation. The discussed method ensures scalability and adaptability by integrating unsupervised clustering with supervised ensemble classification which eventually increases the market segmentation. This research framework helps businesses understand their customers better, thereby targeting the right marketing and proper resource allocation. It has shown a high accuracy of 99.75% and identifies customer segments very effectively (silhouette score: 0.6172).

Keywords: Market Segmentation, Customer Segmentation, Customer Clustering, Ensemble Learning,XGBoost,Gaussian Mixture Model (GMM),Principal Component Analysis (PCA),t-SNE.

P. Ranjan et al. (eds.), *Proceedings of the International Conference on Recent Advances in Artificial Intelligence for Sustainable Development (RAISD 2025)*, Advances in Intelligent Systems Research 196,
https://doi.org/10.2991/978-94-6463-787-8_37

1 Introduction

Market segmentation is regarded as an essential component of every contemporary business plan. By doing this, companies are able to divide their clientele into more manageable, uniform groups according to shared characteristics, habits, or preferences[1],[2]. This gives businesses the chance to customize their marketing plans for more efficient use of resources and higher levels of client satisfaction[3]. With the increasing diversity in consumer needs, there also arises a need for enhanced tools that can make such segmentation more accurate and effective[4]. This paper focuses on the development of a framework for market segmentation using a combination of clustering and classification methods. The goal is to create a system that can identify customers' groups and predict the behavior of those customers in real-time[5]. To accomplish this, data was drawn from real-world sources for a wide range of customer characteristics such as demographics, preferences, and behavioral patterns.
Data preparation and normalization were the first steps while designing methodology[6]. Categorical variables such as gender and product categories were put into numerical formats while continuous data is standardized to maintain consistency within the process. We further applied PCA and t-distributed Stochastic Neighbor Embedding, or t-SNE for simplified analysis[7]. These techniques reduce noise from the data set and give attention to any patterns. While dealing complexity, these methods extract significant information. In terms of clustering, we made use of GMM, and for testing the number of configurations, allows determining the proper number of groupings as well as their clustering parameters[8]. Then to gauge the validity of the obtained clusters silhouette scores of the formed clusters is calculated[9]. We also train an XGBoost classifier to determine which cluster a new customer belongs to[10].
Our system dynamically classifies new customers and finds trends in existing data by combining XGBoost for prediction and GMM for grouping. This combination results in useful and informative to real-world applications while outcomes are also encouraging. The consumer set was successfully classified into the desired cluster with a high classification accuracy by the XGBoost classifier at 99.75%. The silhouette score of high value 0.6172 supports strong clustering quality [11]. As a result, it demonstrates the effectiveness of this framework in producing accurate and trustworthy consumer groups. Using this system, businesses can use it to create more personalized marketing efforts and gain a better understanding of their clients. It can also increase customer loyalty and recommendations by successfully reaching the accurate and perfect segments [12]. Due to its scalable design, which ensures its usage in real-time applications, it is highly flexible in response to changing market conditions.

2 Literature review

Good quality data-driven decision-making forms the essential for effective market segmentation, especially in marketing and customer behavior analysis. An overarching framework of segmentation in stages such as pre-processing, dimension reduction, clustering, and building predictive models has extreme support from existing litera-

ture. This section reviews leading contributions to these domains to contextualize their relevance and applicability.

The quality of input data is the main basis of any machine learning model. According to Sammut and Webb[13], the only way with which the inconsistencies and the biases found in datasets may be counterbalanced is data preprocessing. Encoding methods, which have been used so far for categorical variables are label encoding and one-hot encoding. These transform textual categories into numerical forms where the machine learning model handles them effectively. Han, Kamber, and Pei[14] emphasized the standardization technique such as StandardScaler and MinMaxScaler so that scales of the features may uniformly contribute in the prediction model. It also suggested to convert feature like "TimeSpentPerVisit" into numeric format so that while processing, there won't be any error on account of different formats used [14]. High-dimensional data poses both challenges of computational efficiency as well as interpretability. PCA has been an existing technique with the objective of retaining major information through dimensionality reduction for a long time.

Jolliffe and Cadimaproved that PCA really removes the redundancy among the feature set, retaining as high as 95% variance in data[15]. From the visualization perspective, t-SNE is one of the best techniques to project higher dimensional data into two-dimensional spaces[16]. Studies have shown that t-SNE enhances the visualization of clusters, which means it becomes easier to discern meaningful patterns in complex data sets. In segmentation frameworks, the optimal clusters identification is very essential for actionable insight. GMMs, according to Banfield and Raftery[17], are proven more efficient in dealing with a variety of densities of clusters. This probabilistic model makes GMMs adaptive to diverse distributions of data. Rousseeuw in 1987 first introduced the Silhouette score, currently widely used to evaluate the quality of clustering[18]. It determines an appropriate value for the set of parameters by figuring out whether the data points lie within a compact cluster or are separated from other clusters. More recent studies have used iterative optimization techniques to further fine-tune clustering parameters, such as the number of clusters and covariance types [19].

Visualization is important in interpreting cluster assignments and validating segmentation results. Scatter plots, especially with libraries such as Matplotlib and Seaborn, have been extensively used for visualizing clusters [20]. These tools enable simple definitions of clusters, and most often, these are assisted by color-coding methods that can be used to define the difference between the various clusters. According to Few[21], clear visual information helps the researchers as well as stakeholders make better sense of the data; thus, it becomes part of pragmatic segmentation models. Predictive modeling has a higher relevance with segmentation since new data points into categories can be defined using it. The method is state-of-the-art for such tasks because the gradient-boosting framework of XGBoost excels at handling large, complex datasets [22]. Zhang et al. (2020) showed that analysis of feature importance in XGBoost can be used to draw rich insights into customer characteristics suitable for marketing strategies. Similarly, performance measures like accuracy, precision, recall, and F1-score have become standard measures regarding effectiveness in the predictive models, for example, by Sokolova and Lapalme in 2009. Building dimensionality

reduction, clustering, and predictive modeling techniques into coherent frameworks is fast becoming a developing area.

Wedel and Kamakura [23] discovered in hybrid methods that combine traditional statistical methodology with modern machine-learning methods, in order to enhance scalability and accuracy of segmentation. This is furthered on by Chatterjee [24] through self-supervised learning algorithms with considerable enhancement in coherence in clustering along with prediction accuracy. Yet, despite the progress in this area, challenges have still been found concerning imbalanced data, outliers, and the interpretability of high-dimensional clusters. Hodge and Austin[25] pointed out how data imbalance affects the performance of the model, and to overcome this they proposed several synthetic data generation techniques.

Hamilton [26] then further elucidated that there is so much scope to model such complex relationships by using GNNs to model in data and gave a positive direction to do research work on segmentation [27]. Hybrid Clustering Models and Reinforcement Learning represent other directions for improvement over Segmentation Frameworks. This survey highlights the emergence of market segmentation frameworks where, in fact, preprocessing merges with dimensionality reduction with clustering and predictive modeling at the same time. Methods like PCA and GMMs have formed a backbone, but new methods and techniques like self-supervised learning and GNN are going to bring in something new to this area [28]. These techniques will ensure that scalability, precision, and actionable insights about the decision-making process of data-driven individuals are evolved.

3 Methodology

3.1 Clustering technique-GMM

Clustering techniques are used to generate clusters which are similar to one another based on certain characteristics. The similarity exists is measured by calculating distance the objects within scope. In machine learning, Clustering is one of the unsupervised learning approaches for grouping related objects and occurrences in clusters. Here, Gaussian Mixture Model (GMM) is used. In this work, different number of clusters have been tested which are ranging from 2 to 9. The Silhouette score is used in identifying the optimal combination of cluster numbers and its covariance type. The main reason behind selecting GMM as it perceives data points shall be drawn from mixture of multiple Gaussian distributions.

3.2 Dataset discussion

Every clustering algorithm depends on data supplied to system.Moreover, quality and quantity of available data are most influencing factors. The dataset used here in this work contains customer data aiming to understand customer behavior, and segmentation. The dataset includes combination of categorical features (Gender, Channel, Region, ProductCategory, ProductOwnership and Psychographics etc.) which are then encoded into numerical values which are using Label Encoding while numerical features (TimeSpentPerVisit) which is converted into float values. Here. An new tar-

get variable such as 'Cluster' is created on the basis of clustering results which are representing customer groups. All features excluding Customer Id is normalized by using Standard Scaler for ensuring consistent scale for the purpose of dimensionality reduction and clustering. Principal Component Analysis (PCA) reduced entire dataset into 10 principal components by retaining significant variance. t-SNE projects above components into 2 dimensions for influencing visualization and clustering process. The considered dataset is large enough to perform clustering and classification task. The likely attributes are customer demographics such as gender and region, shopping channels, psychographics and behavioral data.

3.3 Algorithm used

Step 1. The algorithm begins by loading the dataset and preprocessing it, de noted as D

Step 2. Categorical variables Ci∈Dare encoded using Label Encoding: Ci→E(Ci)

Step 3. Feature normalization is applied using the StandardScalerScaler(X) Xnorm=X−μ/σ

Step 4. Dimensionality reduction is conducted using Principal Component Analysis (PCA), retaining k components: XPCA=PCA(Xnorm, k)

Step 5. The reduced data is further transformed into 2D space using t-SNE with perplexity P: Xt-SNE=t-SNE(XPCA,P)

Step 6. Gaussian Mixture Model (GMM) is applied for clustering with n clusters and covariance type GMM(Xt-SNE,n,Σ)→Clusters

Step 7. Optimization of clusters is done by maximizing the silhouette score S: S= b – a/max(a,b)
where a is the average intra-cluster distance, and b is the average inter-cluster distance.

Step 8. The optimal clustering results are visualized in a scatter plot of X t-SNEX with cluster labels.

Step 9. The cluster labels YCluster are used as target labels for supervised learning:
YCluster=GMM(Xt-SNE)

Step 10. A train-test split is performed for classification: (Xtrain,Xtest,Ytrain,Ytest)=Split(Xt-SNE,YCluster)

Step 11. XGBoost classifier is trained on (Xtrain,Ytrain)→Model

Step 12. The model is evaluated using accuracy A and the silhouette score S: A=Correct Predictions/Total Predictions, S=Silhouette Score of Clus ters.

3.4 Classification with XGBoost

In this research work, classification task is implemented by using XGBoost to analyze, evaluate, and utilize customer clusters which are generated during clustering phase. The clustering algorithm which is assigning customer to specific cluster has been treated as target variable for classification task. The dataset was splitted into two

subsets such as training and testing with 80% data for training with XGBoost classifier while rest 20% was reserved for testing process. XGBoost due to its efficiency and accuracy was configured with 500 estimators and maximum tree depth of 10 while learning rate was preserved with 0.05 for optimized performance.

3.5 Flowchart

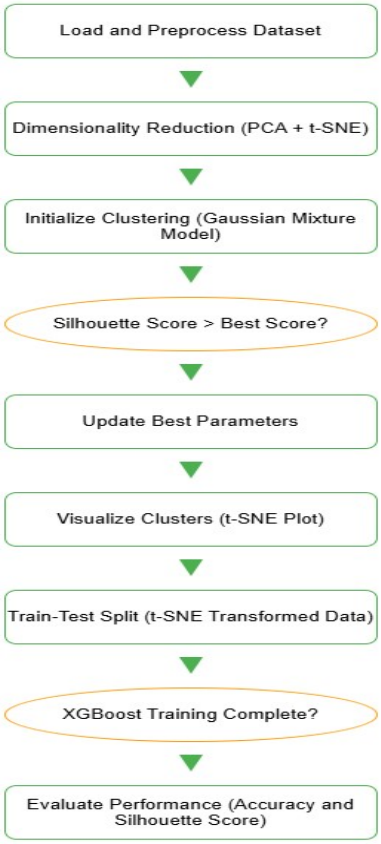

Fig. 1 Flowchart for the Proposed methodology

In **fig. 1**, various stages involved: Load and Process Data Set (The stage involves retrieving unprocessed customer data from various sources and preparing it before analysis. This includes data cleansing, such as addressing missing values, encoding categorial variab es, and scaling numerical features appropriately), Dimensionality Reduction (PCA + t-SNE) includes (Apply dimensionality reduction methods such as PCA to retain essential features and t-SNE for effective visualization for various attributes available), Initialize Clustering -Gaussian Mixture Model (Using GMM to form clusters of customers on the basis of shared characteristics), Evaluating Cluster-

ing Quality -Silhouette Score (Using Silhouette score, quality of clusters is evaluated to measure how well the clsuters are defined and separated), Optimize Clustering Parameters (On the basis of Silhouette score, parameters such as cluster numbers or covariance are modified for better and improved results), Visualizing clusters (Once the formed clusters are optimized, they would be visualized in t-SNE plot providing clear view of groups), Train Test split for classification (for ensuring actionable form of clsuters, t-SNE transformed data is divided into two subsets: traing and testing sets so that next stage of classification can be introduced), Train classification using XGBoost (Using XGBoost as an ensemble learning technique, classification model is built up to keep assigning new customers to identified groups), Model Evaluation (Performance of model is validated by using assess metrics such as Accuracy that states how well the model can classify the customers and Silhouette score for checking how distinct the formed clusters are).

4 Result analysis

The results of this research work reveal the capability of advance ML techniques such as Gaussian Mixture Models (GMM) for the process of clustering and XGBoost for classification task in market/customer segmentation. The Gaussian Mixture Model (GMM) when combined withdimensionality reduction techniques such as PCA and t-SNE are efficiently segmenting the dataset into different customer groups. The efficient clustering optimization which was done here is determined by achieved silhouette score which revealed: Number of clusters: 2, Best Covariance type obswerved: full and Silhouette score is 0.6172.

In fig. 2, clustering results are analyzed by using 2-dimensional t-SNE projection which illustrates two well separately created clusters and are labelled as Cluster 0 in green and Cluster 1 in blue color. The robustness of used clustering methodology can be validated by observing separation and tight cohesion within the clusters where each cluster represents different group of customers with different shared characteristics depicting actionable insights for targeting marketing strategies and analyzing customer behavior.

For clustering, XGBoost classifier was trained to have cluster labels prediction. When classifier was evaluated on test dataset it reveals accuracy: 99.75% and in classification report high precision and recall is observed for both of the clusters. Cluster 0 and Cluster 1 representing unique customer profiles for analyzing targeted campaigns by tailoring product recommendations and personalized services and observed accuracy and silhouette score can be presented as :

- ✓ Accuracy of XGBoost on clustered data: 0.9975
- ✓ Final Silhouette Score for clustering: 0.6172

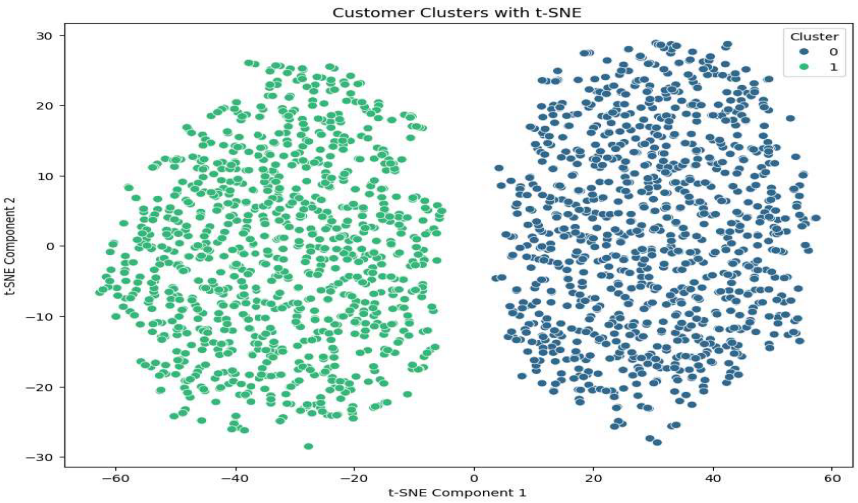

Fig. 2 t-SNE visualization of customer clusters (Cluster 0 and Cluster 1) generated using GMM

CONCLUSION:

This research effectively highlights the application of machine learning methods for successful market segmentation, integrating dimensionality reduction, clustering, and classification. The Gaussian Mixture Model created distinct customer clusters with a silhouette score of 0.6172, indicating the clarity and dependability of the segmentation. The XGBoost classifier provided additional validation to the framework, by attaining significant accuracy of 99.75%, which guarantees accurate identification and categorization of customer segments.

By handling the complexities of original customer data in this approach offers practical insights for a company to refine its marketing strategies, elevate the customer experience, and optimize the product. PCA and t-SNE guaranteed that reducing dimensions maintained essential information, while the inclusion of Gaussian Mixture Models and XGBoost enhanced the framework's robustness.

This efficient and scalable solution offers a reliable tool for organizations aiming to implement data-driven market segmentation.

Disclosure of Interests. The authors declare that there is no conflict of interests

References

1. Shah, B. K.: Sentiments Detection for Amazon Product Review, International Conference on Computer Communication and Informatics (ICCCI) (2021)
2. Kotler, P., Armstrong, G.: Principles of Marketing. Pearson Education (2023).
3. Rust,R.T.,Moorman, C., Zeithaml, V.A.: Customer relationship management: Driving profitable customer interactions, Journal of Marketing Research, 37(3), pp. 310–325 (2000)
4. Berry, M. J. A, Linoff G.: Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management. John Wiley & Sons (2004)
5. Bishop,C. M.: Pattern Recognition and Machine Learning. Springer (2006).
6. Chen,T., Guestrin,C. XGBoost: A scalable tree boosting system, Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining 785–794 (2016)
7. Han, J.Kamber,M.. Pei, J.: Data Mining: Concepts and Techniques. Morgan Kaufmann (2011)
8. Hinton,G. E.: Visualizing data using t-SNE, Journal of Machine Learning Research 9, 2579–2605 (2008).
9. Rousseeuw, P., Silhouettes, J.: A graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics. 20, 53–65 (1987).
10. Gama, J.Žliobaitė, I., Biffl, A., Pechenizkiy, M., Bouchachia, A.: Learning from Data Streams: Adaptation Concepts and Methods. Springer (2014).
11. Reichheld,F. F.: The Loyalty Effect: The Hidden Force Behind Growth, Profits, and Lasting Relationships. Harvard Business School Press (1996).
12. Kedia, V.: Time Efficient iOS Application for Cardiovascular Disease Prediction Using Machine Learning, 5th International Conference on Computing Methodologies and Communication (ICCMC) (2021).
13. Sammut, P., Webb, G.: Encyclopedia of Machine Learning. Springer (2017).
14. Han, J., Kamber,M., Pei,J. : Data Mining: Concepts and Techniques. Morgan Kaufmann (2011).
15. Jolliffe, I. T., Cadima,J.:Principal component analysis: A review and recent developments, Philosophical Transactions of the Royal Society A. 374(2065) (2016).
16. Banfield, J. D., Raftery, A. E.: Model-based Gaussian and non-Gaussian clustering,Biometrics, 49 (3), 803–821 (1993).
17. Rousseeuw,P. J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics, 20, 53–65, 1987.
18. Hastie,T., Tibshirani,R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer (2009).
19. Hunter,J. D.: Matplotlib: A 2D graphics environment, Computing in Science & Engineering 9 (3), 90–95 (2007).
20. Few,S.: Show Me the Numbers: Designing Tables and Graphs to Enlighten. Analytics Press (2012).
21. Chen, T. and Guestrin,C.: XGBoost: A scalable tree boosting system. Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 785–794 (2016).
22. Zhang,H., Tan, C., Wang, C., Wu, X: Enhancing marketing strategies with interpretable machine learning,Journal of Business Research 120, 289–300 (2020).

23. Sokolova,M.,Lapalme,G.: A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45 (4), 427–437 (2009).
24. Wedel, M., Kamakura, W. A.: Market Segmentation: Conceptual and Methodological Foundations. Springer Science & Business Media (2012).
25. Chatterjee, K., Kumar, P., Varshney, B.: Self-supervised learning algorithms for coherent clustering, Advances in Data Science and Machine Learning (2021).
26. Hodge, V. J., Austin, J.: A survey of outlier detection methodologies, Artificial Intelligence Review. 22, 85–126 (2004).
27. Hamilton,W. L., Ying,R., Leskovec, J.: Inductive representation learning on large graphs, Advances in Neural Information Processing Systems 30 (2017).