

The Investigation of Machine Learning Approaches for Customer Segmentation

Ziying Chen^a

Business School, University of Edinburgh, South Bridge, Edinburgh, U.K.

Keywords: Customer Segmentation, Machine Learning, K-Means, Artificial Intelligence.

Abstract: In today's rapidly evolving business landscape, personalized marketing strategies are crucial for businesses to remain competitive. Customer segmentation may assist businesses in understanding distinct clients' needs and preferences, enabling more efficient targeted marketing strategies and customer relationship management. Machine learning techniques serve as key tools for discovering hidden patterns in data. This study aims to give an organised summary of various machine learning algorithms utilized in customer segmentation. The study starts with an overview of machine learning workflow for customer analysis. Then, comprehensive review of the different methods used for customer segmentation in the telecommunication and e-commerce industry is presented based on publications between 2020 and 2024. The study evaluated unsupervised machine learning models, including K-means clustering, Fuzzy c-means, and other more advanced models like K-means++, in terms of their application in customer segmentation. Supervised machine learning models like Random Forest and Artificial Neural Networks (ANN) have also been reviewed. Additionally, the research identifies several limitations in applying machine learning to customer segmentation, including the lack of interpretability, generalizability, and privacy concerns. While possible solutions for integrating technologies such as Expert Systems, Domain Adaptation, and Federated Learning are proposed, which may provide insights on future improvements in machine learning.

1 INTRODUCTION

Customer Segmentation is a predictive analytics technique that enables organizations to categorize customers into distinct groups according to similar behaviors and characteristics. It helps customer base management and creates the foundation for strategic marketing (Kansal et al., 2018). Smith first proposed the idea of customer segmentation in 1956, and the parameters taken into account while segmenting customer can broadly be dependent on the input variables of geographic, demographic, psychographic and behavioral (Das & Nayak, 2022). According to Kansal et al., businesses can better handle their large customer base, understand customers, and conduct customized marketing strategies targeted at diverse customer groups by implementing effective customer segmentation (Kansal et al., 2018). Besides, customer segmentation allows businesses to unravel hidden patterns and trends and better serve their target customers which results in increasing revenues and

enhancing customer segmentation (Kansal et al., 2018).

In the last decade, Machine Learning (ML) has emerged as a powerful tool in many fields as traditional marketing forecasting techniques are failing to perform satisfactory segmentation. It can generate useful results for companies to transform data into information and knowledge with less effort and time. However, several traditional statistical methods and market forecasting approaches, such as the multivariate analysis, are unable to achieve satisfactory segmentation results (Duarte et al., 2022). On the other hand, ML has been proven effective in many different fields of study. For example, ML is employed in the telecom industry. With the advancement of predictive models such as random forest, decision trees, etc., the possibilities to predict customer churn has increased significantly to help businesses retaining their customers (Lalwan , 2022). Furthermore, models that have been presented for detecting credit card fraud (Varmedja et al., 2019) and the prediction of market returns and stock closing

^a <https://orcid.org/0009-0006-3652-3454>

prices (Vijh et al., 2020) demonstrate the precision and effectiveness of machine learning in the financial sector. Most importantly, machine learning models also has been implemented in the field of customer segmentation to solve market-related problems and perform accurate results under vast amount of data.

Besides, there are a great variety of ML techniques that can be utilized for customer segmentation. The most often used is the clustering analysis like K-Means, hierarchical clustering, and Density-based Spatial Clustering of Applications with Noise (DBSCAN). Other improved techniques such as adaptive particle swarm optimization algorithm and combinations with other techniques like Bayes logistic regression are explored by recent innovation studies (Li et al., 2021). Furthermore, there is an increasing number of improvements in the use of customer segmentation with expanding marketing goals and implementation in other fields of study. For instance, clustering can be applied based on other approaches like the Recency, Frequency, and Monetary (RFM) models. While integrating customer segmentation with churn prediction may assist Telecom businesses in predicting and retaining existing customers, as well as understanding churn customers and carrying out retention strategies more precisely (Wu et al., 2021).

Therefore, this paper focuses on providing an overview of customer segmentation using machine learning approaches. The purpose of this study is to provide a thorough investigation into the use of supervised, unsupervised, and other machine learning approaches to produce accurate and reliable customer segmentation solutions. The rest of this article is organized into 3 sections. Section 2 will present several previous studies on customer segmentation using machine learning techniques. Section 3 will discuss the limitations that come with customer segmentation and future directions of this field. Section 4 concludes the investigation and discussion.

2 METHODS

2.1 Introduction of the Machine Learning Workflow

This section will describe the ML pipeline shown in Figure 1 for conducting customer segmentation.

2.1.1 Data Collection

Structured, semi-structured, or unstructured data can be collected from various resources (Zadoo et al., 2022). For customer segmentation problem, data

mainly comes from the company's database (Alkhayrat et al., 2020).

2.1.2 Data Preprocessing

Data is firstly cleaned and analyzed preparing for model training. Missing values, outliers, duplicates, and inconsistencies will be removed or computed using statistical techniques (Goyle et al., 2023). Then, data pre-processing will be carried out for ensuring the data quality.

Data Transformation: For unifying data into a normalized range, data can be encoded using One-Hot encoder or feature scaling techniques like Min-Max-Scaler and Z-Score Standardization method (Goyle et al., 2023).

Dimensionality reduction: Feature selection techniques such as Principal Component Analysis (PCA) are critical for reducing redundant features in the dataset (Alkhayrat et al., 2020), especially in clustering algorithms whose performance degrades with increasing dimension space.

Data Augmentation & Balancing: In some cases where the number of features or instances is insufficient, leading to an imbalanced dataset, undersampling and oversampling methods will be presented to overcome this challenge (Frye et al., 2021).

2.1.3 Model Architecture

Depending on the project goal, a variety of supervised, semi-supervised, and unsupervised ML techniques might be selected.

2.1.4 Model Training

Models are trained utilizing features obtained in previous phases and training data split from the original dataset.

2.1.5 Model Testing

Testing approaches vary depending on the model being used. Area Under the Curve (AUC), recall, F-measure, and precision are often used by supervised machine learning algorithms to assess performance. While clustering methods are frequently examined using the silhouette coefficient (Alkhayrat et al., 2020).

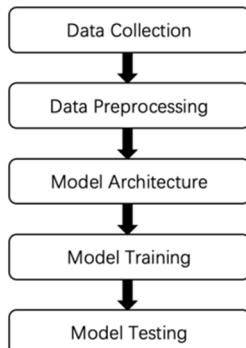


Figure 1: The workflow of developing the machine learning algorithm (Photo/Picture credit: Original).

2.2 Telecommunication Industry

2.2.1 Unsupervised Machine Learning Algorithm

K-means clustering has been proven to be one of the most popular unsupervised machine learning techniques for partitioning database into K distinct clusters (Sharaf Addin et al., 2022). The desired number of clusters is predetermined. The K-means algorithm assumes that all data points fall into a single cluster and categorize unlabelled data points to one of the k centroids closest to the given point measured by Euclidean distance as the initial cluster assignment. The clustering process is repeated and further improved until each data point has the least variation from its centroid (Shen, 2021).

Sharaf Addin et al. utilized behavioral patterns to segment customers and provide more information about their preferences and average usage of telecom services. As a result, the paper identified 4 clusters that group telecom customers' mobile behavioral data into different labeled groups (Sharaf Addin et al., 2022). Furthermore, Wu et.al. proposed an integrated customer analytics approach for churn management by conducting both churn prediction and customer segmentation (Wu et al., 2021). Churn prediction is first proposed, and the study conducted customer segmentation focusing on only the churn customers. Customers are divided into 3 clusters using K-Means clustering, and each segment's characteristics is categorized with the probability of churning and demand level.

2.2.2 Supervised Machine Learning Algorithm

Businesses can manage customer churn by employing supervised ML models for customer segmentation and classification. ML algorithms are used in the

telecom sector to classify customers with the label of churn and non-churn customer. While applying supervised ML algorithms of artificial neural networks (ANN) and linear discriminant analysis (LDA), Mahmoud formed several clusters based on the customers' usage, behavior, and loyalty (Mahmoud & Asyhari, 2024). On the other hand, a hybrid technique that combines the chi-squared automatic interaction detector (CHAID) decision tree technique and K-means clustering was proposed by Pejić Bach et al. Initially, the dataset has been grouped into clusters using K-means clustering (Pejić Bach et al., 2021). Furthermore, classification models are created using decision trees to discover churn determinants, which assist in comprehending the distinctness of characteristics and behaviors various customer categories.

2.3 E-Commerce Industry

2.3.1 Unsupervised Machine Learning Algorithm

Databases of E-commerce platform is mainly unlabeled, so there are many unsupervised ML algorithms to construct customer segmentation. In the study done by Shen (Shen, 2021), RFM analytic model is used to segment customers from a real-world ecommerce online transaction database. Firstly, RFM values are derived during the feature engineering phase. Then the Elbow Method is utilized to calculate the optimal number of clusters for K-means clustering. Clusters with more than 300 customers' RFM scores are examined for gaining information on customers' behaviors and provide recommendations for personalized strategies. Furthermore, unique product descriptions are counted using the Term Frequency-Inverse Document Frequency (TF-IDF) method and grouped into different categories using the K-means clustering algorithm. The product and customer categories are further linked using Apriori Algorithm's association rules mining.

Further research conducted by Snehalatha et al., utilized DBSCAN and Fuzzy c-means clustering techniques to partition the UK e-commerce dataset. Without specifying the optimal number of clusters, the DBSCAN algorithm can divide the datasets into numerous distinct bunches. It performs well with vast amounts of data even when noise and outliers are present. Besides, the fuzzy C-means algorithm may group a data point into multiple clusters applying the membership function, and the iteration minimizes the objective function of the distance weighted by the membership (Snehalatha et al., 2023).

While addressing the limitation of ML algorithms in capturing dynamic changes in customer data,

Sivaguru proposed a modified dynamic fuzzy c -means (dFCM) clustering technique. This advanced algorithm can update the customer segmentation system with new information to discover changes in the customer segments for the dynamic customer segmentation of a retail supermarket (Sivaguru, 2022). Based on the RFM patterns, the dFCM clustering updated the cluster solution from four initial clusters into five different segments, and the supermarket may target appropriate marketing strategies based on the customers' time to time change to gain more profit and customer satisfaction.

In another study conducted by Jing (Jing, 2024), the AFCH model was designed to assist e-commerce businesses in implementing targeted marketing and customer retention strategies. The study expands on the traditional customer value matrix AF, where A (average purchased amount) represents customer contribution, and F (purchase frequency) reflects customer satisfaction. It introduces two additional variables: total clicks C, which indicates consumer attention, and customer hold time H, which represents customer loyalty. The study applies three clustering algorithms: K-means, Self-Organizing-Map (SOM) + K-means and K-means++ to segment the data. In the SOM + K-means approach, data is first processed through the SOM network, where the weighting vectors of the SOM neurons are continuously modified in a competitive layer until the iterations are attained. The resulting cluster centers and number of clusters are then implemented in the K-means algorithm to achieve the final clustering results. In the K-means++ algorithm, an initial centroid is randomly selected, and the distance from each data point to this initial centroid is calculated. The next centroid is determined by calculating the probability of each sample being chosen as the next cluster center using the roulette wheel selection method, which then feeds into the K-means algorithm. Since centroids are relatively dispersed, the K-means++ method yields the most accurate clustering results with the smallest sum of squared errors (SSE). The SOM + K-means algorithm also improves clustering accuracy by overcoming the K-means algorithm's limitation of subjectively selecting the number of clusters. The article ultimately presents K-means++ clustering results by categorizing customers into four groups of iron, copper, silver, and gold customers based on their AFCH values, providing insights into e-commerce customer value.

2.3.2 Supervised Machine Learning Algorithm

Arefin proposed an extensive analysis of consumer behaviors using the UK retail dataset which combines

K-means clustering based on RFM model and multiple supervised machine learning models. This includes Logistic Regression, Random Forest, AdaBoost, ExtraTrees, LGBM, and XGBoost which are used to predict customer lifetime value segments. Then, the performance is compared using the metrics of accuracy, precision, recall, F-measure, and AUC curve (Arefin, 2024).

3 DISCUSSIONS

3.1 The Discussion Related to the Limitations of Machine Learning Algorithms

Although ML algorithms are vital for making customer segmentation, determining the accuracy of these algorithms can be challenging. Therefore, it is critical to understand ML limitations during practical business applications.

A major limitation of machine learning, particularly deep learning, is its lack of interpretability. Complex black-box models, such as neural networks and K-means, are difficult to explain to stakeholders since they provide no information about how they reach their conclusions or the underlying cognitive processes. While the outcomes of these models can be helpful, if the model is unable to elucidate the factors contributing to consumer behavior segmentation, it is potentially problematic for businesses to trust the results. For example, if an algorithm classifies a client as likely to churn without explaining to the business why this client is classified as a churn customer rather than a retain customer, stakeholders may find it difficult to develop strategies based on these conclusions without understanding the rationale behind the results.

Secondly, the lack of generalizability across datasets or industries remains a critical limitation for businesses using machine learning for customer segmentation. Traditional machine learning segmentation models trained on historical data may not be suitable for future customer data, which may change due to market trends, seasonal variations, or socio-economic factors. Consequently, models that perform well on past data might fail to generalize accurate segmentation results when facing new and unseen datasets. Furthermore, customer data might vary widely across regions, demographics, or industries, making it difficult for a pre-trained model based on a specific region or industry to effectively adapt to multiple tasks. Similarly, Farahani et al. observed that machine learning algorithms often experience performance degradation when training and test data have different distributions in real-world

applications (Farahani et al., 2021). Therefore, this lack of generalizability might lead to suboptimal marketing strategies and increased costs as businesses need to develop different models for customers from various regions or demographics.

Thirdly, the heavily dependence of customer segmentation machine learning algorithms on substantial volumes of customer data raises significant privacy concerns. Machine learning algorithms rely on the quality and availability of data, which frequently includes customers' purchase history, browsing behaviors, and personal information. The collection and utilization of these data pose serious privacy issues to both businesses and consumers. Therefore, ensuring the privacy and security of customer data from misuse or leakage while leveraging it for effective customer segmentation presents considerable challenges for businesses.

3.2 Future Prospects

3.2.1 The Possible Solutions for Interpretability

As the use of data becomes more complicated, researchers should develop more advanced technologies to improve models' efficiency and accuracy. Interpretable models hold great promise for balancing accuracy and interpretability in black-box machine learning models. Tan et al. suggested the utilization of Expert Systems to assess the significance of their explanation capability (Tan et al., 2016). The Expert Systems may analyze their reasoning processes and explain their decisions made by extracting knowledges from human experts for purposes such as interpretation, prediction, and classification.

Additionally, the SHapley Additive exPlanations (SHAP) method can derive intricate details and explanations about model decisions by revealing the significance of each variable and its contribution to the model output. SHAP values can rank feature importance from most to least significant, helping stakeholders in identifying the most crucial factors influencing customer segmentation or prediction values (Luo et al., 2023). This knowledge also enhances targeted marketing activities toward each segment by adding insights into the distinctions between different segmented customers. Furthermore, businesses can validate these results by verifying whether these parameters are consistent with real-world scenarios, and therefore enhancing the reliability and applicability of machine learning models.

3.2.2 the Possible Solutions for Generalizability

Addressing the challenge of generalizability, researchers may incorporate transfer learning and domain adaptation into customer segmentation. Transfer learning trains dataset by reusing models that have been trained on different but related source domains. This technique helps to improve the performance on new and unseen data while lessen the reliance on extensive target domain data. By leveraging knowledge from diverse datasets, such as those from other related product lines or demographics, transfer learning can enhance the efficiency of machine learning applications when facing substantial variations in customer data. For business operating across multiple regions, domain adaptation method can help businesses adjusting to different customer behaviors more effectively by aligning the feature distributions between training and target domains. This ensures the model trained in one region can be accurately applied in other geographies region, enhancing the model's robustness and adaptability.

3.2.3 The Possible Solutions for Privacy

To mitigate privacy concerns, future research may focus on privacy-preserving machine learning algorithms. Federated learning, for example, allows models to be trained through decentralized approaches. This technique enables the model to learn from a diverse range of data while keeping the raw data on client devices. By avoiding the sharing and transferring of raw training data with any third-party entity or central servers, federated learning may prevent the potential data breaches and enhance the data security.

4 CONCLUSIONS

Based on publications from 2020, this paper presents a systematic overview of machine learning algorithms used in the telecommunication and e-commerce industry for customer segmentation. Machine learning approaches have been divided into supervised and unsupervised models. K-means clustering has been identified as the most suitable and frequently used technique for conducting customer segmentation. Additionally, other advanced ML algorithms are introduced for their applications for segmenting customers, including Fuzzy C-means, Artificial Neural Networks (ANN), and K-means++.

Furthermore, this study analyzes the limitations of ML algorithms in real-world scenarios, particularly related to interpretability, generalizability, and data privacy. To improve the practical application of machine learning in customer segmentation, future research might focus on improving interpretability through Expert Systems and SHAP, enhancing generalizability through Transfer Learning and Domain Adaptation, and addressing privacy concerns through Federated Learning.

REFERENCES

- Alkhayrat, M., Aljnidi, M., & Aljoumaa, K. 2020. A comparative dimensionality reduction study in telecom customer segmentation using deep learning and PCA. *Journal of Big Data*, 7(1), 9.
- Arefin, S., Parvez, R., Ahmed, T., Ahsan, M., Sumaiya, F., Jahan, F., & Hasan, M. 2024. Retail Industry Analytics: Unraveling Consumer Behavior through RFM Segmentation and Machine Learning. In *24th Annual IEEE International Conference on Electro Information Technology (eit2024)*.
- Das, S., & Nayak, J. 2022. Customer segmentation via data mining techniques: state-of-the-art review. *Computational Intelligence in Data Mining: Proceedings of ICCIDM 2021*, 489-507.
- Duarte, V., Zuniga-Jara, S., & Contreras, S. 2022. Machine learning and marketing: A systematic literature review. *IEEE Access*, 10, 93273-93288.
- Farahani, A., Voghoei, S., Rasheed, K., & Arabnia, H. R. 2021. A brief review of domain adaptation. *Advances in data science and Information engineering: proceedings from ICDATA 2020 and IKE 2020*, 877-894.
- Frye, M., et al. 2021. Benchmarking of data preprocessing methods for machine learning-applications in production. *Procedia CIRP*, 104, 50-55.
- Goyle, K., Xie, Q., & Goyle, V. 2024. Dataassist: A machine learning approach to data cleaning and preparation. In *Intelligent Systems Conference* (pp. 476-486). Cham: Springer Nature Switzerland.
- Jing, Z. 2024. Research on E-commerce Customer Segmentation Based on the K-means++ algorithm. In *International Conference on Advanced Information Networking and Applications* (pp. 439-446). Cham: Springer Nature Switzerland.
- Kansal, T., Bahuguna, S., Singh, V., & Choudhury, T. 2018. Customer segmentation using K-means clustering. In *2018 international conference on computational techniques, electronics and mechanical systems (CTEMS)* (pp. 135-139). IEEE.
- Lalwani, P., Mishra, M. K., Chadha, J. S., & Sethi, P. 2022. Customer churn prediction system: a machine learning approach. *Computing*, 104(2), 271-294.
- Li, Y., Chu, X., Tian, D., Feng, J., & Mu, W. 2021. Customer segmentation using K-means clustering and the adaptive particle swarm optimization algorithm. *Applied Soft Computing*, 113, 107924.
- Luo, Y., Zhang, R., Wang, F., & Wei, T. 2023. Customer Segment Classification Prediction in the Australian Retail Based on Machine Learning Algorithms. In *Proceedings of the 2023 4th International Conference on Machine Learning and Computer Application* (pp. 498-503).
- Mahmoud, H. H., & Asyhari, A. T. 2024. Customer Segmentation for Telecommunication Using Machine Learning. In *International Conference on Knowledge Science, Engineering and Management* (pp. 144-154). Singapore: Springer Nature Singapore.
- Pejić Bach, M., Pivar, J., & Jaković, B. 2021. Churn management in telecommunications: Hybrid approach using cluster analysis and decision trees. *Journal of Risk and Financial Management*, 14(11), 544.
- Sharaf Addin, E. H., Admodisastro, N., Mohd Ashri, S. N. S., Kamaruddin, A., & Chong, Y. C. 2022. Customer mobile behavioral segmentation and analysis in telecom using machine learning. *Applied Artificial Intelligence*, 36(1), 2009223.
- Shen, B. 2021. E-commerce customer segmentation via unsupervised machine learning. In *The 2nd international conference on computing and data science* (pp. 1-7).
- Sivaguru, M. 2023. Dynamic customer segmentation: a case study using the modified dynamic fuzzy c-means clustering algorithm. *Granular Computing*, 8(2), 345-360.
- Snehalatha, N., et al. 2023. Customer Segmentation and Profiling For E-Commerce Using Dbscan And Fuzzy C-Means. *Proceedings on Engineering*, 5(3), 539-544.
- Tan, C. F., Wahidin, L. S., Khalil, S. N., Tamaldin, N., Hu, J., & Rauterberg, G. W. M. 2016. The application of expert system: A review of research and applications. *ARPN Journal of Engineering and Applied Sciences*, 11(4), 2448-2453.
- Varmedja, D., Karanovic, M., Sladojevic, S., Arsenovic, M., & Anderla, A. 2019. Credit card fraud detection-machine learning methods. In *2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH)* (pp. 1-5). IEEE.
- Vijh, M., Chandola, D., Tikkiwal, V. A., & Kumar, A. 2020. Stock closing price prediction using machine learning techniques. *Procedia computer science*, 167, 599-606.
- Wu, S., Yau, W. C., Ong, T. S., & Chong, S. C. 2021. Integrated churn prediction and customer segmentation framework for telco business. *Ieee Access*, 9, 62118-62136.
- Wu, S., Yau, W. C., Ong, T. S., & Chong, S. C. 2021. Integrated churn prediction and customer segmentation framework for telco business. *Ieee Access*, 9, 62118-62136.
- Zadoo, A., Jagtap, T., Khule, N., Kedari, A., & Khedkar, S. 2022. A review on churn prediction and customer segmentation using machine learning. In *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-ITCON)* (Vol. 1, pp. 174-178). IEEE.