K-MEANS CLUSTERING APPROACH BASED INTELLIGENT CUSTOMER SEGMENTATION TO INCREASE SALES USING CUSTOMER PURCHASE BEHAVIOR DATA

[1]P. Vishal, [2]Dr.G.Venkata Koti Reddy

[1]M.Tech Scholar, Department of Computer Science and Engineering, Holy Mary Institute Of Technology and Science, Bogaram (V), Keesara (M), Hyderabad, Telangana, India
[2]Professor & HOD, Department of Computer Science and Engineering, Holy Mary Institute Of Technology and Science, Bogaram (V), Keesara (M), Hyderabad, Telangana, India
Corresponding Email : gvkotiteddy@gmail.com

ABSTRACT: Currently, e-commerce platforms are increasingly used and can be found in almost every business. An e-commerce system is a platform for internet marketing and product promotion. The separation of a target based on shared qualities is known as segmentation, which is the process of putting together groups of consumers. Customer segmentation is utilized to choose how to communicate with each sort of consumer in order to maximize each customer's profit to the increase. With the use of a customer segmentation research, marketers may effectively categorize different consumer types based on demographic, behavioral, and other factors by making use of the large amounts of information on both present and potential customers. This analysis describes a K-Means Clustering technique based Intelligent Customer Segmentation to increase sales using Customer Purchase Behavior Data. The K-means clustering is a learning method that is used to analyze the acquired data and segment the clients. Mean Squared Error (MSE) and coefficient of determination (R2) are two metrics they use to assess the prediction model. For a high number of observations, K-Means clustering performs properly.

KEYWORDS: Purchase behavior, K-Means clustering, customer segmentation, Cluster analysis.

DOI: 10.48047/ecb/2023.12.7.350

I. INTRODUCTION

Due to of the accessibility of extensive historical data and the resulting rise in economic competitiveness, and it is now normal practice to use data mining techniques to find critical and strategic information hidden in organizational data.[1].

The data mining is the process of taking logical information from a dataset and presenting it in an easy-to-understand style for decision assistance. Many data analysis activities may now be solved using modern intelligent information technology to increase corporate productivity and decision making [2]. They can analyze huge quantities of data from multiple sources using intelligent technology, detect trends, and predict economic performance. These activities include international economic activity research predictions, inventory management, client basket analysis, and etc.[3].

Customers may become confused when they receive too much information or unwanted data that are unrelated to their frequent purchases or their interest in the items. As a result, their customers could decide not to buy the things they needed, and the company might lose some potential clients. The clustering analysis will be useful in classifying E-commerce clients based on their spending and purchasing patterns as well as any particular brands or products they may be interested [4]. Customer segmentation is one of the most effective methods used in business analytics to analyze and categorize consumer behavior. Techniques for clustering are used by customers with equivalent methods, ends, and behaviors are put together into homogenous groups [5].

Customer segmentation supports organizations in finding or revealing separate groups of consumers that think and behave differently, as well as exhibit different patterns of consumption [6]. Clustering algorithms identify groupings that are dissimilar internally but similar externally [7]. Processing the use of clustering techniques, target marketing efforts may be carried out more successfully by identifying unique client types and dividing the consumer base into groups of people with similar profiles. Customers differ in terms of their actions, needs, desires, and behavioral characteristics.

A statistical method called highlight clustering is quite similar to categorization. It aggregates reasonably uniform observations and significant groupings of raw data. The items in a specific cluster share some qualities and characteristics with those of other clusters, but they are distinct from one another [8]. The categorization is completed by comparing data similarities based on features identified in raw data. The major goal was to determine the perfect amount of clusters. Hierarchical and non-hierarchical clustering techniques are the two main categories. The process of clustering is an in progress, information may be extracted iteratively from large amounts of unstructured raw data. To achieve the best results for a certain classification task, an appropriate clustering technique and parameters must be selected. Exploratory data mining techniques like clustering are utilized in a number of application-oriented fields like machine learning, classification, and pattern identification [9]. Data mining is now increasing significantly for knowledge-based services like distributed and grid computing.

With the use of a customer segmentation plan, businesses may target certain consumer characteristics are makes it easier to allocate marketing resources more effectively and raises the chance of cross- and up-selling [10]. Businesses find it easier to provide original offers to get customers to spend more when they deliver customized communications to a group of consumers as part of a marketing mix matched to their requirements [11]. Through enhancing interactions with clients, consumer segmentation can be helpful in retaining customers and increasing customer loyalty. Given that they are more personalized than impersonal brand communications that disregard purchase history or any other kind of customer relationship, The consumer support marketing materials that utilize customer segmentation more after receiving them. [12].

This has been shown that clustering helps customer segmentation. In unlabeled datasets, clustering a type of unsupervised learning allows us to find clusters. The remaining work is arranged as follows: The study is concluded with Section V, which follows Section II's detailed explanation of the literature review and Section III's explanation of the methodology.

II. LITERATURE SURVEY

X. Chen, W. Sun, B. Wang, Z. Li, X. Wang and Y. Ye, et. al. [13] provides a set of level weights to differentiate the weights of various tree levels and a set of sparse node weights to differentiate the weights of various tree nodes in a buy tree to evaluate the dissimilarity of two consumers expressed through two purchase trees. This proposal describes a PurTree (Purchase Tree) subspace metric to measure this dissimilarity. Six alternative clustering techniques were compared with Two-level subspace weighting (TSW) spectral clustering using ten benchmark data sets. The results of the experiments demonstrate the new method's superiority. Chen X., Yang M., Fang Y., Nie F., Zhao Z. and Huang J. Z., et. al. [14] utilizing the "personalized product tree," also known as the "purchase tree," to symbolize a customer's transaction history. Consequently, the set of client transaction data may be compressed into a collection of purchases trees. For quick clustering of purchase trees, they suggest the partitional clustering technique is named “PurTreeClust”. By allocating each client to the closest representation, the clustering results are then obtained. Ten real-world transaction data sets were used in a series of examinations, and the results demonstrate the proposed method's higher performance.

D. Maryani, A. Astuti, R. D. Riana, Ishaq, Sutrisno and E. A. Pratama, et. al. [15] identifies the use of data mining techniques based on the RFM (Recency, Frequency, and Monetary) model and clustering approaches to categorize clients for Nine Reload Credit. K-Means is the name of the clustering method used. Additionally, they used the K-Means method to a cluster analysis, with the results showing that there were 39 and 63 customers in cluster 2 respectively. The results of this study may be utilized by the firm to determine the types of customers, then it will understand how to maintain consumer ownership.

R. Taniguchi, Y. Ohtaka and S. Morishita, et. al. [16] Cellular Automata (CA) and a simulation were used to simulate client purchase behaviour in a business. Each client's transaction data, often known as "Point of Sales (PoS) data has been captured. These data enable retailers to identify popular or attractive products for each customer. Local neighbor rules in the CA algorithm are described as the interaction of the phenomenon component parts. They developed a model that portrays the steps consumers take when they consider their alternatives for planned and unforeseen purchases. The consumer passed about the shop as a result of the product's placement. Zhu J., M. Zhu, Wang H., B. Tsou K. and M. Ma, et. al. [17] evaluates unlabeled, unstructured, free-form textual customer feedback without providing any questions in order to carry out aspect-based opinion research. A multi-aspect bootstrapping approach is initially devised in order to discover a number of factors that are connected to each aspect and are utilized to identify each aspect. Second, a model for segmenting sentences into basic single-aspect components using an aspect-based method is suggested for polling purposes. The last step is the detailed presentation of an aspect-based opinion polling method. Real-world Chinese restaurant evaluations implemented in experiments showed that described method can perform aspect-based opinion polling tasks with an accuracy of 75.5%.

X. Zhang, G. Feng and H. Hui, et. al. [18] examines the special characteristics of the Customer Relationship Management (CRM) system used by the telecommunications sector and proposes a customer churn model based on segmenting customers. Customers are initially divided into groups, and the features of the high value customer group are identified using the improved Fuzzy C-means clustering method. Secondly, through the use of SAS **(**Statistical Analysis System**)** data mining technology, a prediction model of customer-churn is implemented utilizing historical data and SAS Enterprise Miner. The outcome of customer segmentation is then put to a customer-churn model in order to obtain an accurate list of lost customers. The efficiency of this strategy in lowering customer churn is demonstrated through an experiment. T. Jiang and A. Tuzhilin, et. al. [19] In order to create customer segments, they provide a direct grouping-based approach that divides consumers into groups based on how much transactional data from various customers is combined, as opposed to computed statistics, and then uses that information to create data mining models of consumer behaviour for every group. The next step is to construct client segments by utilizing combinatorial optimization to determine the best way to categorize the customer base. It is demonstrated that, even if it is computationally tractable, the optimal direct grouping method continually outperforms the statistics-based and one-to-one approaches in the majority of experimental conditions. T. Iwata, K. Saito and T. Yamada, et. al. [20] Describe a unique recommendation technique that can be used for both measured and subscription services are maximizes the possibility that Life Time Value (LTV) will be increased. Using maximum entropy models, they deduce a user's interests from their purchasing history and use those interests to enhance suggestion. Since improved customer satisfaction leads to higher LTV, Customers and online stores benefit from described approach. Two sets of real log data for measured and subscription services are used to assess described technique.

III. INTELLIGENT CUSTOMER SEGMENTATION

The work flow of K-Means Clustering technique based Intelligent Customer Segmentation to increase sales using Customer Purchase Behavior Data is shown in Fig. 1.

Collecting data on E-commerce purchases is the first stage. Next, determine whether the data shows a clustering pattern. For the machine learning analysis of customer purchase patterns used in this analysis, the Malaysian Digital Economy Corporation (MDEC) repository's Malaysian E-commerce dataset was implemented. There are 285,000,000 consumer purchase histories in this project's dataset, which includes E-Commerce behavior data from a multi-category business.

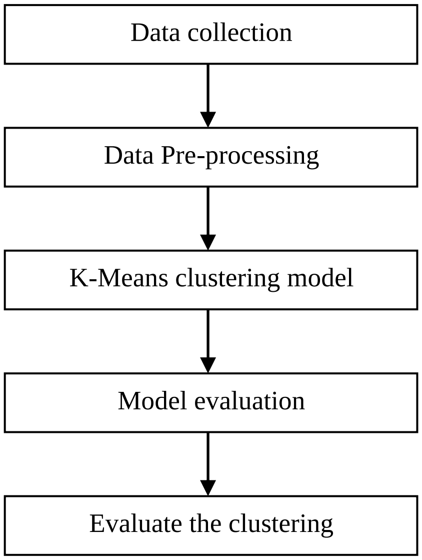

Fig. 1: WORK FLOW OF INTELLIGENT CUSTOMER SEGMENTATION MODEL

The questionnaire response will be reviewed, and an analysis based on the gathered quantitative data will be produced. The questionnaire's questions are regarded as the research's basis since they will provide a statistical analysis utilizing the data gathered. Before running the algorithm on the data sets selected for this study, pre-processing methods will be used to detect missing values, inaccurate information, and other anomalies. RStudio and Microsoft Office Excel will be the tools used for the K-Means technique-based data pre-processing.

Several clusters, including c1, c2, c3, and cn, will then be created from the dataset, using the K-Means technique. Unsupervised learning, or K-Means clustering, is a technique used to address clustering-related issues. The process of categorizing a dataset into a predetermined number of clusters using K-Means clustering involves k centers to each cluster. Since different places give different effects, the k-centers should be carefully positioned. Every cluster should be as widely apart as possible for the best results. The optimal cluster size, denoted as k, which the dataset's greatest distance can be accurately determined. Elbow technique is one method for determining the ideal number of clusters. The best run would be chosen using a specified criterion after comparing the outcomes of several runs with multiple k. Generally speaking, a large k increases the risk of overfitting while reducing inaccuracy.

Examine the clustering outcomes after describing each class in terms of one or more rules based on its own characteristics and those of the data products that make up that class.

Mean Squared Error (MSE) and coefficient of determination (R^2) are two metrics they use to evaluate the mentioned models. The suggested approach is compared to Support Vector regression (SVR) and Means-multiple Linear Regression (MLR), previously developed techniques to show the reliability. The clustering outcome is validated for the real application if it is very reliable. Using a new clustering method if necessary, the clustering analysis is repeated. In summary, the K-Means algorithm increases efficiency for companies that perform online sales by enhancing the quality of client data clustering.

IV. RESULT ANALYSIS

Based on the evaluations of the inputs, consumers were clustered using the k-means clusters. For the machine learning analysis of consumer purchase behaviors in this analysis, the Malaysian E-commerce dataset from the MDEC repository was used. Training and test sets were created from the data in each cluster. The prediction models were built using training sets, which made up 60% of the total data, and tested against test sets, which made up 40% of the total data. Mean Squared Error (MSE) and Coefficient of Determination (R^2) are two metrics they use to assess the provided model. End users may see the outcomes of customer segmentation due to data visualization and dashboards.

The dashboard's overall view is seen in Fig. 2. In the image below, a summary of each e-category code, brand, cost, item description, and event type is displayed. On this page, the data distribution and overall analysis for all the examined characteristics or variables.

Fig. 2: OVERVIEW OF DASHBOARD

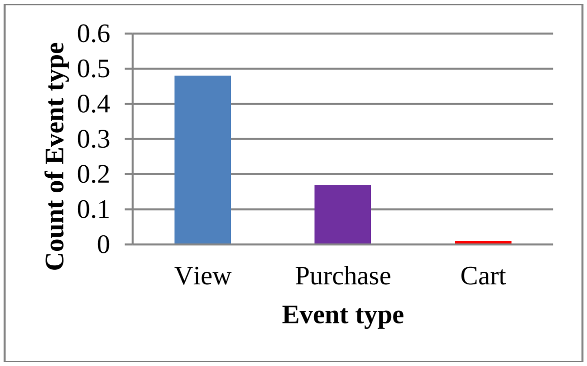

Fig. 3: EVENT TYPE DASHBOARD

The customer segmentation dashboard's product brand page is seen in Fig. 3 below.

This paper provides information about the brand that online shoppers approval. From the most popular brand of products to the least popular product, the visualizations display this information. Additionally, the dashboard displays a table that lists how many product brands each user has browsed, bought, or put to their cart for purchase.

An estimator's Mean Squared Error (MSE) quantifies the average of the squares of the mistakes in a method for estimating an unobserved variables, or the square root of the mean difference between the actual the estimated values. The predicted value of the squared error loss corresponds to the risk function known as MSE. MSE is almost never precisely positive (and never negative), this is the outcome of chance or the estimator's disregard for information that may have produced a more precise estimate.

The performance of a statistical model in estimating a result is shown by the coefficient of determination (R^2). The dependent variable in the model is a representation of the result. R^2 can have a value of 0 or 1, with 1 being the highest possible number.

Comparisons are conducted between the new technique and the earlier approaches, Support vector regression (SVR) and Means-multiple linear regression (MLR) to demonstrate the suggested method's robustness.

Table 1: COMPARATIVE PERFORMANCE ANALYSIS

Method	**Mean Squared Error (MSE)**	**coefficient of determination (R^2)**
Support Vector Regression (SVR)	0.16	0.86
Means-Multiple Linear Regression (MLR)	0.18	0.88
K-Means Clustering	0.09	0.98

The Fig. 4 and Fig. 5 show the graphical representation of Mean Squared Error (MSE) and coefficient of determination (R^2) parameters for described K-Means Clustering approach, Means-Multiple Linear Regression (MLR) and Support Vector Regression (SVR) models.

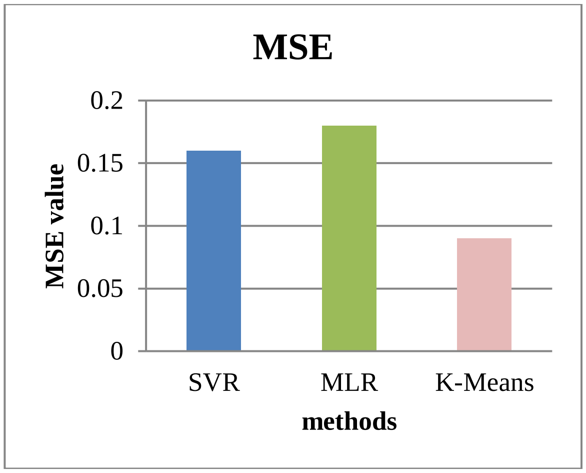

Fig. 4: COMPARATIVE ANALYSIS IN TERMS OF 'MSE'

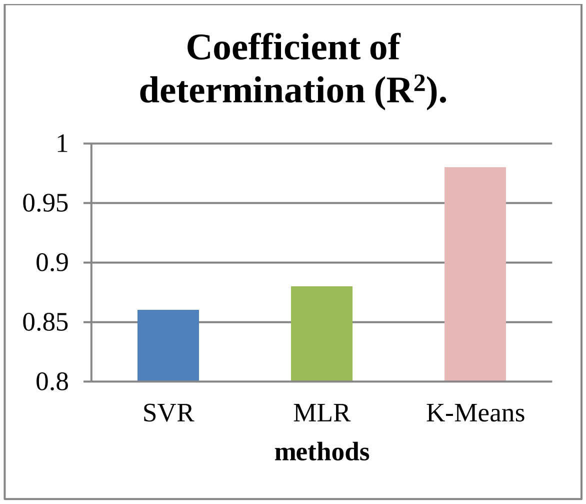

Fig. 5: COMPARATIVE ANALYSIS IN TERMS OF 'COEFFICIENT OF DETERMINATION (R^2)'

With a reduced MSE and a better Coefficient of Determination (R^2), the results show that K-Means Clustering has enhanced the performance of Intelligent Customer Segmentation. When vendors are faced with volatility in data analysis, As a result of the present research's results and analysis, providers are better able to interact with customers since they can use that knowledge to guide marketing choices.

V. CONCLUSION

K-Means Clustering Approach Based Intelligent Customer Segmentation To Increase Sales Using Customer Purchase Behavior Data is described in this analysis. Through targeting specific consumer groups using a customer segmentation plan, businesses may allocate marketing resources more effectively, increasing the chance of cross- and up-selling. The machine learning analysis of customer purchase patterns in the malaysian E-commerce dataset from the MDEC repository was used for this investigation. A learning method is used with K-Means clustering to evaluate the given data and segment the clients. The process of categorizing a dataset into a predetermined number of clusters using K-Means clustering assigning k centers to each cluster. The outcomes of client segmentation are presented to end users through data visualization and dashboards. Mean Squared Error (MSE) and Coefficient of Determination (R^2) are two metrics they use to determine the provided model. Results indicate that the implement of K-Means Clustering has enhance the performances of Intelligent Customer Segmentation with lower MSE and higher Coefficient of determination (R^2). When faced with a volatile data analysis, providers may engage customers more successfully by utilizing the results of the current study and interpretation to guide marketing choices regarding customers e-commerce buying behaviors. This project have done with as minimum flaws as possible and can further be enhanced by including major identification of statistics of people and improving the accuracy of the output. This project implemented k-means algorithm, it can be further enhanced by using few complex algorithms such as conventional neural networks algorithms.

VI. REFERENCES

[1] Z. -H. Sun and X. Ming, "Multicriteria Decision-Making Framework for Supplier Selection: A Customer Community-Driven Approach," in IEEE Transactions on Engineering Management, vol. 70, no. 10, pp. 3434-3450, Oct. 2023, doi: 10.1109/TEM.2021.3089279.

[2] Z. Yang, Q. Li, V. Charles, B. Xu and S. Gupta, "Online Product Decision Support Using Sentiment Analysis and Fuzzy Cloud-Based Multi-Criteria Model Through Multiple E-Commerce Platforms," in IEEE Transactions on Fuzzy Systems, doi: 10.1109/TFUZZ.2023.3269741.

[3] M. Zavali, E. Lacka and J. de Smedt, "Shopping Hard or Hardly Shopping: Revealing Consumer Segments Using Clickstream Data," in IEEE Transactions on Engineering Management, vol. 70, no. 4, pp. 1353-1364, April 2023, doi: 10.1109/TEM.2021.3070069.

[4] Q. Jiang and Y. Jiang, "Analysis of e-commerce customer data mining based on Apriori optimization algorithm," 2022 2nd International Signal Processing, Communications and Engineering Management Conference (ISPCEM), Montreal, ON, Canada, 2022, pp. 155-160, doi: 10.1109/ISPCEM57418.2022.00037.

[5] M. Ghahramani, A. O'Hagan, M. Zhou and J. Sweeney, "Intelligent Geodemographic Clustering Based on Neural Network and Particle Swarm Optimization," in IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 52, no. 6, pp. 3746-3756, June 2022, doi: 10.1109/TSMC.2021.3072357.

[6] A. Solichin and G. Wibowo, "Customer Segmentation Based on Recency Frequency Monetary (RFM) and User Event Tracking (UET) Using K-Means Algorithm," 2022 IEEE 8th Information Technology International Seminar (ITIS), Surabaya, Indonesia, 2022, pp. 257-262, doi: 10.1109/ITIS57155.2022.10009981.

[7] B. Melović, B. Rondović, S. Mitrović-Veljković, S. B. Očovaj and M. Dabić, "Electronic Customer Relationship Management Assimilation in Southeastern European Companies—Cluster Analysis," in IEEE Transactions on Engineering Management, vol. 69, no. 4, pp. 1081-1100, Aug. 2022, doi: 10.1109/TEM.2020.2972532.

[8] S. Miloudi, Y. Wang and W. Ding, "A Gradient-Based Clustering for Multi-Database Mining," in IEEE Access, vol. 9, pp. 11144-11172, 2021, doi: 10.1109/ACCESS.2021.3050404.

[9] P. Zhou, C. Lu, J. Feng, Z. Lin and S. Yan, "Tensor Low-Rank Representation for Data Recovery and Clustering," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 5, pp. 1718-1732, 1 May 2021, doi: 10.1109/TPAMI.2019.2954874.

[10] Y. Yuan, K. Dehghanpour, F. Bu and Z. Wang, "A Data-Driven Customer Segmentation Strategy Based on Contribution to System Peak Demand," in IEEE Transactions on Power Systems, vol. 35, no. 5, pp. 4026-4035, Sept. 2020, doi: 10.1109/TPWRS.2020.2979943.

[11] A. Syaputra, Zulkarnain and E. Laoh, "Customer Segmentation on Returned Product Customers Using Time Series Clustering Analysis," 2020 International Conference on ICT for Smart Society (ICISS), Bandung, Indonesia, 2020, pp. 1-5, doi: 10.1109/ICISS50791.2020.9307575.

[12] Y. Feng, X. Wang and L. Li, "The Application Research of Customer Segmentation Model in Bank Financial Marketing," 2019 2nd International Conference on Safety Produce Informatization (IICSPI), Chongqing, China, 2019, pp. 564-569, doi: 10.1109/IICSPI48186.2019.9095900.

[13] X. Chen, W. Sun, B. Wang, Z. Li, X. Wang and Y. Ye, "Spectral Clustering of Customer Transaction Data With a Two-Level Subspace Weighting Method," in IEEE Transactions on Cybernetics, vol. 49, no. 9, pp. 3230-3241, Sept. 2019, doi: 10.1109/TCYB.2018.2836804.

[14] X. Chen, Y. Fang, M. Yang, F. Nie, Z. Zhao and J. Z. Huang, "PurTreeClust: A Clustering Algorithm for Customer Segmentation from Massive Customer Transaction Data," in IEEE Transactions on Knowledge and Data Engineering, vol. 30, no. 3, pp. 559-572, 1 March 2018, doi: 10.1109/TKDE.2017.2763620.

[15] I. Maryani, D. Riana, R. D. Astuti, A. Ishaq, Sutrisno and E. A. Pratama, "Customer Segmentation based on RFM model and Clustering Techniques With K-Means Algorithm," 2018 Third International Conference on Informatics and Computing (ICIC), Palembang, Indonesia, 2018, pp. 1-6, doi: 10.1109/IAC.2018.8780570.

[16] R. Taniguchi, Y. Ohtaka and S. Morishita, "Prediction of Purchase Behavior of Customers in a Store by Cellular Automata," 2015 Third International Symposium on Computing and Networking (CANDAR), Sapporo, Japan, 2015, pp. 436-441, doi: 10.1109/CANDAR.2015.37.

[17] J. Zhu, H. Wang, M. Zhu, B. K. Tsou and M. Ma, "Aspect-Based Opinion Polling from Customer Reviews," in IEEE Transactions on Affective Computing, vol. 2, no. 1, pp. 37-49, Jan.-June 2011, doi: 10.1109/T-AFFC.2011.2.

[18] X. Zhang, G. Feng and H. Hui, "Customer-Churn Research Based on Customer Segmentation," 2009 International Conference on Electronic Commerce and

Business Intelligence, Beijing, China, 2009, pp. 443-446, doi: 10.1109/ECBI.2009.86.

[19] T. Jiang and A. Tuzhilin, "Improving Personalization Solutions through Optimal Segmentation of Customer Bases," in IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 3, pp. 305-320, March 2009, doi: 10.1109/TKDE.2008.163.

[20] T. Iwata, K. Saito and T. Yamada, "Recommendation Method for Improving Customer Lifetime Value," in IEEE Transactions on Knowledge and Data Engineering, vol. 20, no. 9, pp. 1254-1263, Sept. 2008, doi: 10.1109/TKDE.2008.55.