International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 05 | May – 2025 DOI: 10.55041/ISJEM03836
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Role of Data Fusion on Customer Profiling and their Lifetime in Retail Sector and it's Performance Evaluation.

Nainsi[1], Priyavrat Raghuvanshi[2], VaibhavTyagi[3], Rohit Kumar Singh[4]

[1,2,3]Department of Information Technology
Meerut Institute of Engineering and Technology, Meerut UP, India Pin-250005
nainsi.krishan.itl.2021@miet.ac.in,
priyavrat.raghuvanshi.itl.2021@miet.ac.in vaibhav.tyagi.itl.2021@miet.ac.in
[4]Department of Electronics & Communication Engineering
Meerut Institute of Engineering and Technology, Meerut UP, India Pin-250005 rohit.singh@miet.ac.in

Abstract- Being profitable was a crucial goal for the banking sector's long-term pros- perity, and building solid, enduring connections with customers is largely dependent on their level of satisfaction. Through the analysis of buyer information, companies could offer customized services. Banks may improve customer experiences, personalize their product lines, and locate growth possibilities through the use of data analysis methods. The accomplishment of managing client relationships (CRM), which permits ongoing client development and retention, depends on customer classification and profiles. By employing these tactics, banks can increase their clientele, offer personalized products, make money, and maximize cross-selling and up-selling campaigns.

Keywords*:* *K-means clustering, hierarchical clustering, data mining, customer profiles, lifetime value and RFM analysis.*

1 Introduction

Understanding consumer behaviour and preferences is crucial for business expansion in the cutthroat retail industry. This project looks at how data fusion enhances lifecycle value estimation and customer segmentation in the retail sector. It investigates how combining various data sources can:

- Improve consumer profiles' accuracy and comprehensiveness.
- Make prediction models for lifetime value estimation stronger. Enhance mar- keting tactics by gaining deeper understanding of consumer behaviour.

The study illustrates the significance of data fusing in modern retail method by evaluating its effects in different domains and provides useful advice for compa- nies looking to optimize customer value.

The Purpose and Goals of the study: Purpose:

- The purpose of this investigation is to investigate how data fusion might im- prove customer profiling and life time value (LTV) forecasting within the re- tail sector.
- Its goal is to assess whether applying data fusion techniques might enhance the lifetime value (LTV) estimations as well as customer profiling.

Goals:

- To identify key data sources used in customer profiling within the retail sector.
- To explore the role of data fusion in integrating customer data from diverse sources.

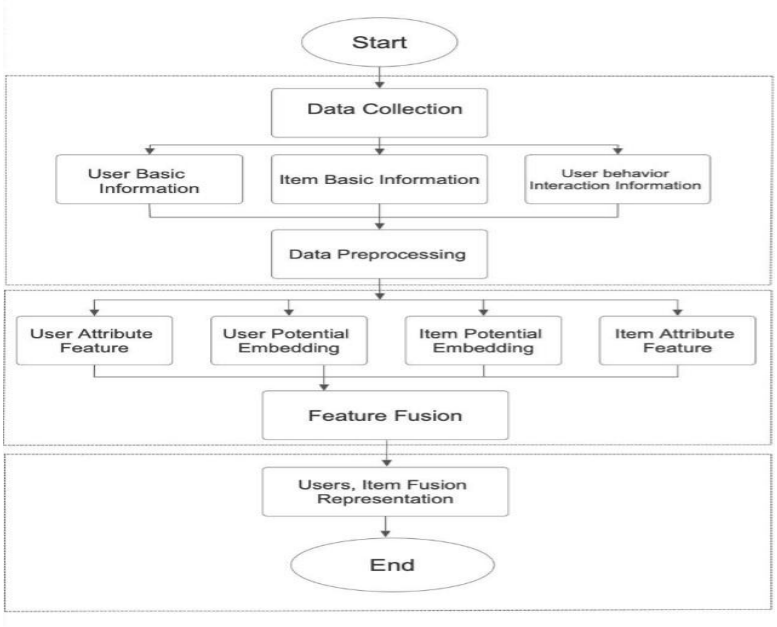

Fig. 1: Data Fusion Diagram

2. Literature Review

Customer relationship management (CRM) comprises a set of processes and enabling systems supporting a business strategy to build long-term, profitable relationships with specific customers. Customer data and information technology (IT) tools form the foun- dation upon which any successful CRM strategy is built. [1]. In today's competitive market, businesses use data mining to identify potential customers and enhance direct marketing. This study explores a preprocessing method for customer profiling using RFM analysis and a boosting tree for prediction, improving sales performance [2]. Re- cent advancements in smart point-of-sale (POS) systems have revolutionized offline retail by enabling businesses to collect and analyze consumer data, such as purchase history. This study presents a customer profiling method that integrates coarse-grained value classification with fine-grained purchase behavior prediction [3]. Since the 1980s, Customer Relationship Management (CRM) has become vital for boosting profitability and competitiveness through strong customer connections. This paper enhances the tra- ditional RFM model by introducing an incremental RFM approach, improving customer segmentation. [4]. Predicting customer intentions is crucial for increasing sales, with technologies like machine learning and IoT offering valuable solutions. This paper examines an IoT-based system using cameras and image processing to predict preferences and recommend products in stores, particularly for undecided shoppers. [5]. Personalized marketing is crucial for retaining online customers, and customer segmen- tation is a key strategy. A common technique for figuring out the ideal number of clus- ters is the Elbow Method, improving segmentation efficacy [6]. This study explores how CRM, or customer relationship management, might help SMEs by combining cus- tomer information management with innovation promotion. The evaluation emphasizes the benefits of Customer Relationship Management and (CKM) combined, illustrating that the pair increases business efficiency through better decision-making and an in- creased awareness of how consumers behave [7]. CRM

technologies belong for max- imizing satisfaction with clients, which is a key component of business success. The following paper explores data mining techniques in customer relationship management for client retention & churn control, with a focus on classification using the K-Means technique enabling targeted advertising [8]. The internet has changed everyday life, with e-commerce becoming a fast-growing industry. These reviews not only influence purchasing decisions but also provide valuable insights for businesses to refine their marketing strategies and improve products and services [9]. The K-means clustering algorithm is an unsupervised learning method that groups data into K clusters. The final result minimizes the total distance, or loss function, between points and their centers. The loss function can be calculated using the following formula:

$$J = \sum_{i=1}^{K} \sum_{x \in C} d(x, \mu_i)^2 \quad (1)$$

The Barabasz and Kalinski index and elbow method are used to determine the optimal number of clusters in this study. It analyzes two datasets: Member Information (194,760 entries) and Sales Transactions (1,893,532 entries), covering data from January 1, 2015, to January 3, 2018 [10]. Customer Relationship Management (CRM) leverages tech- nology to gather, manage, and analyze data from multiple customer touchpoints. This approach helps businesses foster long-term relationships, personalize marketing, and adapt products and services to meet customer needs, ultimately gaining a competitive advantage [11]. Customer Relationship Management (CRM) originated in the 1990s and has become central to business strategies. It emphasizes the continuous process of initiating, maintaining, and terminating customer relationships, rather than being just a technology or one-time strategy [12]. Developing a customer segmentation model using the K-Means clustering algorithm on sales data enables precise customer grouping, which aids in targeted marketing and enhances customer relationship management. This research seeks to improve segmentation accuracy by utilizing big data techniques, enhancing business decision-making and boosting customer satisfaction, loyalty, and profits [13]. Clustering is essential in data mining and market segmentation. This study uses K-Means clustering and the SPSS Tool to predict supermarket sales across sea- sonal cycles, enabling data-driven decision-making [14]. Customer segmentation helps businesses understand customer needs and make informed decisions. Introduced by Smith in 1956, it classifies customers based on demographics, behavior, and geography. [15]. Associating, correlating, and merging data and information from one or more sources is known as the fusion of data. Enhancing position and identity estimations and facilitating thorough and quick situational and danger evaluations are its mainobjectives. [16]. Customer segmentation involves two key stages: Foundation, where business goals and strategies are aligned, and Analysis, where customer value is eval- uated. The Synthesis stage helps create a profile of the ideal audience by understanding their preferences and behaviors, enabling marketers to tailor their strategies [17]. In today's business world, customer segmentation is crucial for companies to meet their clients' varied needs. By grouping customers based on characteristics like de- mographics, behaviors, and purchasing patterns, businesses can customize their mar- keting strategies. [18]. Purchases or jobs are significantly impacted by the UK retail market, and as Big Data has grown, data-driven decision-making has become essential for businesses, particularly those in the retail industry [19]. Businesses are depending upon the data analysis in modern rapid online marketplace to understand consumer be- haviour and obtain a competitive advantage. Businesses are using sophisticated meth- ods like data mining, machine learning, and artificial intelligence to examine customer preferences, buying habits, and interaction patterns [20]. Market segmentation divides a market into distinct groups based on needs, behaviors, or characteristics, requiring tailored marketing strategies. Segmentation can be based on factors like geography, demographics, and behaviors, with segments evaluated for stability, growth, and align- ment with company goals [21]. Customer Relationship Management (CRM) is a strat- egy aimed at building long-term customer relationships to enhance satisfaction and loy- alty. It helps businesses analyze customer value, target high-value segments, and in- crease profits [22].

3. Dataset Description:

Arthur Hughes developed the RFM concept in 1994, measures customer behaviour based on Recency, Frequency, and Monetary factors. It helps businesses recognize high-value customers, improve retention, and optimize marketing strategies.

Table 1: Advantages and disadvantages of previous traditional algorithms and machine learning algorithms in the field of customer segmentation research.

Serial No.	Research Methods	Research object	Research limitations
Literature1 [45]	Probability models: Pa- reto/NBD and BG/NBD Ma- chine learning algorithms: generalized	Retailer data during large online promo- tions	The overall prediction result of machine learning algorithm is lower

	additive model and support vector machine		than the actual mean value.
Literature2 [46]	K-means clustering method, K-medoids method and Fuzzy RFM model	Online sales transac- tion data	there are only three categories of customer segmentation.
Literature3 [47]	Fuzzy c-means cluster and the RFM model	Customer consumption data of e-commerce platform	According to the char- acteristics of custom- ers, the fuzzy c-means cluster and the RFM model is used to seg- ment customers, but the customer lifecycle value is calculated.
Literature4 [48]	Tree clustering, RFM model	Customer transaction data	This research combines Tree clustering with

			RFM model, and only conducts customer seg- mentation research.
Literature5 [49]	RFM model, K-Means and Fuzzy C-Means algorithms	Online retail data	This study only calcu- lates the number of cus- tomer clusters accord- ing to customer charac- teristics.

Data mining (KDD) uncovers valuable patterns from large datasets through stages like data collection, preprocessing, and knowledge extraction. Recency, Frequency, & Monetary (RFM) measures are utilized in this study for classifying the online retail customers. To successfully carry out the classification process, the investigation uses a methodical strategy that consists of key stages:

- Data Sources:

The Online Retail II dataset from a UK-based online retailer having a large bulk clientele that specializes in unusual presents is used in this study. Nine essential attributes, including region, number of cases, and donation date, are included in the dataset.

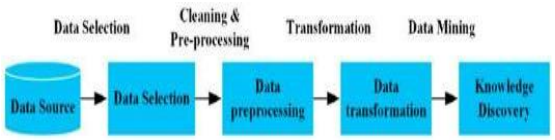

Fig2: KDD Process Flow

- Data Selection:

The dataframe.info () method, which provides information on attribute names, data types, and the quantity of non-null numbers, is employed in this paper for summarizing the set of data using the language's panda's module. Each among the parameters in the dataset, such as "Invoice No," "Stock Code," and "Description," is described in relation to the study.

- Data Preprocessing:

This study preprocesses information using Python's pandas by ad Dressing missing values with Drona () and eliminating duplicate using the duplicated () and drop duplicates () routines. Additionally, outliers disappear, such as can- celled orders identified by invoice numbers beginning with "C".

- Data Transformation:

Recency (time since last purchase), Frequency (number of orders), and Mon- etary (total expenditure) were the three factors used in this study to calculate RFM values.

- Knowledge Discovery:

The R-, F-, and M-scores are used to segment customers; high-value custom- ers perform better than average in each of these categories. Eight segmentation combinations are produced as a result, including "key customers" and "churned customers."

4. Model Visualization:

By minimizing the sum of squared distances between points of data or their centroids, the K-Means clustering method separates information into K clusters. K-Means is used to categorize customers based on factors including spending score, income, and age. By identifying the "elbow" point in a Within-Cluster Sum of Squares (WCSS) plot, the Elbow Method assists in determining the ideal number of clusters in Fig.3.

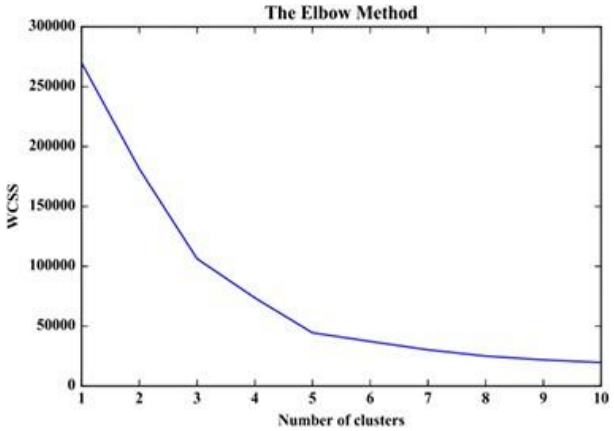

Fig.3: Finding the optimal number of clusters using Elbow method for K Means Clustering

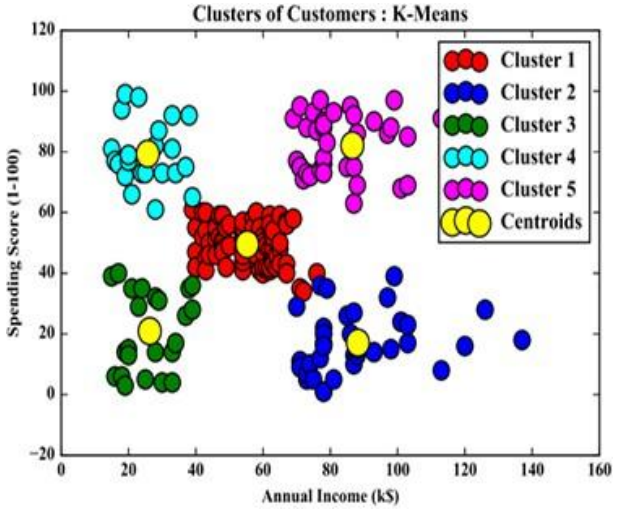

Fig.4: Clusters formed as a result of applying K-Means Clustering on the dataset taken for study

When n is the number of data points, the k is the number of clusters, and l is the number of iterations, the algorithm's temporal complexity is O(nil). O (k + n) is a space-time complexity. Hierarchical clustering uses either a divided (top-down) and agglomerative (from the bottom up) technique to produce a network of groups. Each data point begins simply an individual group in Agglomerative Hierarchical Grouping (AHC), and clus- ters are progressively joined based on similarity until a single cluster is created. Divi- sive Hierarchical Clustering (DHC) divides a single cluster into smaller ones iteratively in Fig. 4.

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 05 | May – 2025 DOI: 10.55041/ISJEM03836
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

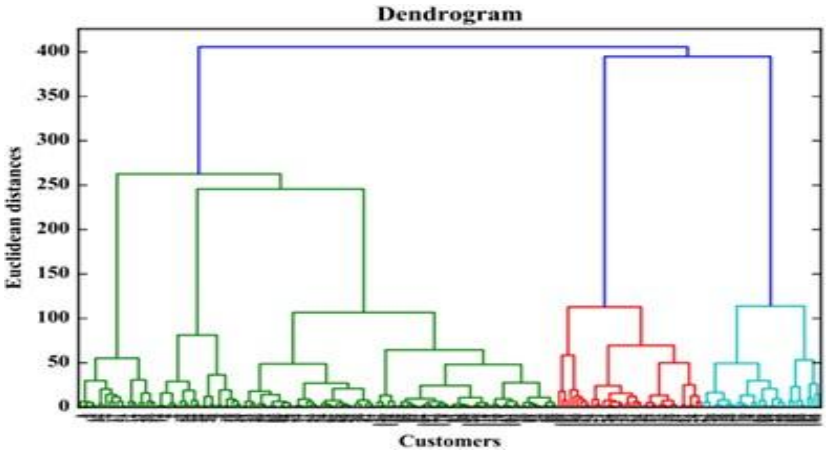

Fig. 5: Visualization of the formation of clusters in the studied dataset with the help of a dendrogram.
The optimal number of clusters in hierarchical clustering is determined by placing a cut-off line at a specific

threshold on the dendrogram.

Fig.6: Clusters formed as a result of applying Hierarchical Clustering on the dataset taken for study
A scatter plot with clustering groups data points based on similarities in Annual Income (X-axis) and Spending Score (Y-axis) in Fig.6.

- Clusters: Clusters in a scatter plot group individuals with similar Annual In- come and Spending Scores, with different colors representing distinct seg- ments.
- Data points: Data points correspond to individuals, positioned based on their income and spending behavior.

5. Result and Discussion

The passage compares K-Means and Hierarchical Clustering for customer segmentation, highlighting key differences in their strengths and limitations, particularly for large datasets:

5.1 K-Means vs. Hierarchical Clustering: Summary

1. **Computational Efficiency**:
   - K-Means is fast and scalable, ideal for large datasets, but requires predefining the number of clusters (K), which can influence results.
   - Hierarchical Clustering is computationally intensive for large datasets but doesn't require predefined clusters and offers a flexible hierarchical structure.
2. **Suitability for Data Size**:
   - K-Means efficiently handles large datasets, as it operates iteratively and scales well.
   - Hierarchical Clustering works better with smaller datasets, as recalculating the proximity matrix adds overhead but provides deeper insights through dendro- grams.
3. **Handling Random and Small Datasets**:
   - K-Means can be impacted by the starting position of centroids, which makes it sensitive to randomness.
   - Hierarchical Clustering works well for smaller datasets and provides flexibil- ity in choosing the number of clusters by using dendrograms after the analysis.
4. **Cluster Structure**:
   - K-Means generates flat, non-hierarchical clusters, which might not capture the complexities of the relationships in the data.
   - Hierarchical Clustering produces a dendrogram, giving a deeper insight into the relationships within the data and the process of cluster formation.
5. **Accuracy and Performance**:
   - K-Means works well with larger datasets but may suffer from reduced accu- racy if the wrong K is chosen or if the data is poorly separated.
   - Hierarchical Clustering generally provides better accuracy, especially with smaller or more complex datasets.

5.1.1 Hybrid Approach:

By combining both techniques, K-Means can be used for initial segmentation because of its speed, while Hierarchical Clustering can refine the results to generate more pre- cise and detailed clusters.

5.2 Customer classification based on the RFM Model:

The RFM model measures customer value through three factors: Recency (time since the last purchase), Frequency (how often purchases are made), and Monetary (overall spending).

Table 2: Dataset-related information introduction.

Dataset char- acteristics	Multivariate, Order, Time Series, Text	Number of in- stances	541909	Region	Business
Attribute characteristi cs	Integer, Real	Number of at- tributes	8	Donation date	2019-09-21
Related tasks	Classificatio n, Regression, Clustering	Lack of value	Yes	Network hits	106790

Table 3: Attribute information of raw data.

Attribute	Information	Meaning
Invoice No	Invoice number	A 6-digit integral number uniquely assigned to each transac- tion. If this code starts with the letter ‘c’, it indicates a can- celation.
Stock Code	Product (item) code	A 5-digit integral number uniquely assigned to each distinct product.
Description	Product (item) name	Description of product name
Quantity	Numeric	The quantities of each product (item) per transaction.
Invoice Date	Invoice date and time	The day and time when a transaction was generated.
Unit Price	Unit Price	Product price per unit in sterling (ˆ A£).
Customer ID	Customer number	A 5-digit integral number uniquely assigned to each cus- tomer.
Country	Country name	The name of the country where a customer resides.

In this study, the "Customer ID," "Invoice Date," "Unit Price," and "Quantity" columns are used as target variables. The "Invoice Date," "Unit Price," and "Quantity" are cho- sen to build the feature data for the RFM model, providing a comprehensive analysis of customer behavior.

Fig 7: Missingness map.

Table 4: Customer classification and customer characteristics.

Customer type	R value	F value	M value
Important value cus- tomer	High	High	High
Important development customer	High	Low	High
Important protection customer	Low	High	High
Important retention customer	Low	Low	High
General value cus- tomer	High	High	Low
General development customer	High	Low	Low
General retention cus- tomer	Low	High	Low
Lost customer	Low	Low	Low

Figure 7 depicts the distribution of customers and their spending habits, according to the customer classification shown in Table 4. Key engagement approaches include con- verting low-value customers to high-value, strengthening relationships with retained users, and maintaining loyalty among high-value customers to sustain revenue.

Table 5 Operational strategies for different customer groups.

Customer type	Behavior characteristics	Operation strategy
Important value customer	Recently, this customer group has purchased, with high pur- chase frequency and high con- sumption, and they are the main consumers.	Upgrade to the very important person (VIP) customers, provide personalized services, and tilt more resources.
Important development cus- tomer	Recently, this customer group has purchased, with low pur- chase frequency and high cus- tomer unit price.	Provide member points service and provide a certain degree of discount to improve the retention rate of customers.
Important protection customer	Recently, this customer group has not bought, but the purchase frequency is high and the con- sumption is high.	Introduce the latest prod- ucts/functions/upgraded services through SMS and email to pro- mote customer

		consumption.
Important retention customer	Recently, this customer group has not bought, and the purchase frequency is low, but the cus- tomer unit price is high.	Introduce the latest prod- ucts/functions/upgrade services, promotional discounts, etc., through SMS, email, phone, etc., to avoid the loss of customers.
General value customer	Recently, this customer group has purchased, with high pur- chase frequency, but low con- sumption.	Introduce the latest prod- ucts/functions/upgraded services to promote customers' consump- tion.
General development customer	Recently, this customer group has purchased, with low pur- chase frequency and low con- sumption.	Provide community services, in- troduce new products/functions, and promote customers' con- sumption.
General retention customer	Recently, this customer group has not bought, with high pur- chase frequency and low con- sumption.	Introduce new products/func- tions to arouse this part of cus- tomers.
Lost customer	Recently, this customer group has not bought, with low pur- chase frequency and low con- sumption, which has been lost.	This part of customers can be aroused by promotion and dis- count. When the resource alloca- tion is insufficient, this part of users can be temporarily aban- doned.

Table 6: The AUC value and accuracy value benchmarking table of nine algorithms.

Classifier	Accuracy	AUC
K-nearest neighbors	90.25%	74.65%
Logistic regression	91.36%	86.14%
Support vector machine	91.14%	67.25%
Decision tree	87.14%	68.67%
Random forest	89.98%	73.94%
AdaBoost	94.60%	93.82%
Gradient boosting decision tree	93.80%	95.12%
Naive Bayes	88.18%	85.71%
Multilayer perceptron	91.14%	87.17%

The evaluation results in Table 6 show that boosting-based fusion algorithms, like Ada- Boost and Gradient Boosting Decision Trees (GBDT), deliver the highest AUC and accuracy values, highlighting their superior performance compared to other methods.

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 05 | May – 2025 DOI: 10.55041/ISJEM03836
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

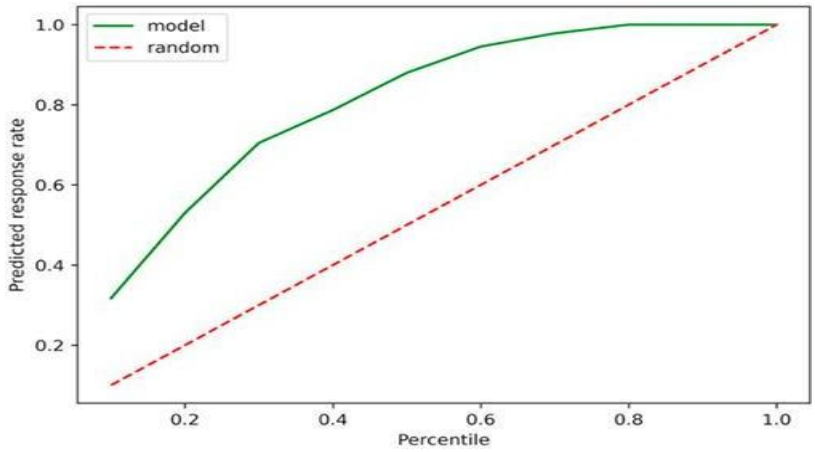

Fig8: Yield Curve

Table7: Value and confidence interval distribution of model parameters.

	Coeff	Se (Coeff)	Lower95% bound	Upper 95% bound
α	9.30268544	0.337551659	8.641084188	9.964286691
r	1.784611036	0.052432477	1.681843382	1.887378691
a	0.059187585	0.006403364	0.046636992	0.071738178
b	1.132869909	0.081850595	0.972442743	1.293297074

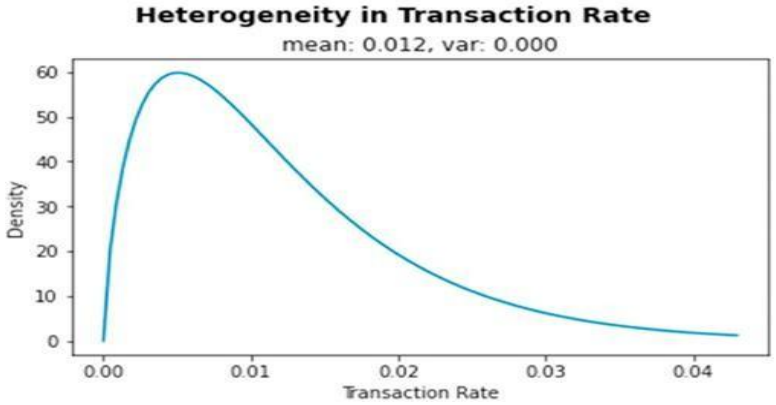

Fig9: Customer's purchase tendency curve

Conclusion and Future Scope

The K-means clustering method segments department store customers into distinct groups, providing insights into their behaviors and values. Using metrics like the Ba- rabasz and Kalinski index and the elbow method, four unique customer segments are identified. These profiles help create personalized marketing strategies and integrate online and offline shopping experiences, improving the overall customer journey.

References

[1] Ngai, E. W. T., Xiu, L., & Chau, D. C. K. (2009). Application of data mining tech- niques in customer relationship management: A literature review and classification. *Ex- pert Systems with Applications*, *36*(2), 2592-2602. https://doi.org/10.1016/j.eswa.2008.02.021.

[2] "Attention-based models for speech recognition," by J. Chronowski, D. Bandana, D. Serdyuk, K. Cho, and Y. Bengio, arrive preprint arXiv:1506.07503 (2015).

[3] [An RFM model adaptable to product catalogues and marketing criteria using fuzzy linguistic models: case study of a retail business], Mathematics, vol. 9, no. 16, p. 1836, 2021; R. G. Martínez, R. A. Carrasco, C. Sanchez-Figueroa, and D. Gavilan.

[4] M. Mohan, M.W. Nyadzayo, R. Casidy, Customer identification: the missing link between relationship quality and supplier performance, Ind. Market. Manag. 97 (2021) 220–232.

[5] Carvajal-Trujillo, E.; Cabrera-Sánchez, J.-P.; Ramos-de-Luna, I.; Villarino-Ramos, Á.F. Online Recommendation Systems: Elements Affecting Their Use in E-Commerce. 2020, Sustainability 12, 8888.

[6] Sukru Ozan, "A Case Study on Customer Segmentation by using Machine Learn- ing Methods", IEEE, Year: 2018.

[7] Fischer, J.; Leventon, J.; Newegg, J.; Schomer us, T.; Valmeyer, U.; Abson, D. J., … N. W. Jager (2017). Make use of points to transform sustainability. Aggarwal, S. (1997). Ambo, 46(1), 10-30. doi:10.1007/ s13280-016-0800-y.

[8] Utilizing an operational study technique, Adebiyi SO, Olaoyes, & (2016) en- hanced customer churn and retention decision management. 6(2):12–21 EMAJ Emerg Mark J

[9] "A Collective Data Mining Approach to Predict Customer actions," Third Interna- tional Conference on Intelligent Communication Technologies and Virtual Mobile Net- works (ICICV), 2021, E. Manohar, P. Jenifer, M. S. Nisha, and B. Benita.

[10] Ahmed M. Seraj R, Islam S.M.S. (2020). The k-means algorithm: A comprehen- sive survey and performance evaluation. Electronics, 9(8):1295. https://doi.org/10.3390/electronics9081295.

[11] Abu Ghazaleh, M. and Zabdi A.M. (2020), "Promoting a revamped CRM through internet of things and big data: an AHP based evaluation", International Jour- nal of Organizational Analysis, Vol. 28 No.1, pp. 66-91.

[12] B. Libia, Y. Bart, S. Gensler, C.F. Hofacker, A. Kaplan, K. Katterheinrich, E.B. Kroll, Brave new world? On AI and the management of customer relationships, J. In- teract. Market. 51 (1) (2020) 44–56, https://doi.org/10.1016/j.intmar.2020.04.002.

[13] Anitha, P., & Patil, M. M. (2022). RFM model for customer purchase behavior using K-Means algorithm. Journal of King Saud University-Computer and Infor- mation Sciences, 34(5), 1785–1792.

[14] I. S. Dhillon and D. M. Modha, "Concept decompositions for large sparse text data using clustering," Machine Learning, vol. 42, issue 1, pp. 143-175, 2001.

[15] Najiha A, Radfar R, Malayali S S. Data mining application for customer segmen- tation based on loyalty: An Iranian food industry case study. 2011 IEEE International Conference on Industrial Engineering and Engineering Management. IEEE, 2011: 504
- 508.

[16] Boström, H., Andler, S. F., Brodhead, M., Johansson, R., Karlsson, A., van La- ree, J., Niklasson, L., Nilsson, M., Persson, A., & Ziemke, T. (2007). *On the defini- tion of information fusion as a field of research.* Institutional för Communication ouch Information, Skived, Sweden.

[17] Tushar Kansal, Suraj Bahuguna, Vishal Singh, Tanupriya Choudhury "Customer Segmentation using K-means Clustering", IEEE, Year: 2018.

[18] Kumar, V.; Reinartz, W. Customer Relationship Management; Springer: Ber- lin/Heidelberg, Germany, 2018.

[19] Lekh war, S., Yadav, S., & Singh, A**.** (2019). Big data analytics in retail. In *In- formation and Communication Technology for Intelligent Systems: Proceedings of ICTIS 2018, Volume 2* (pp. 469–477). Springer, Singapore.

[20] Ode Dina, C**.** (2023). *Impact of big data on marketing strategy and consumer be- haviour analysis in the US.* SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4520361.

[21] Agresti, A. (2002). Categorical data analysis. Hoboken, New Jersey: Wiley.

[22] E. Ngai, L. Xiu and D. Chau, "Application of data mining techniques in customer relationship management: A literature review and classification", Expert Systems with Applications, vol. 36, no. 2, pp. 2592-2602, 2009