Customer Data Analytics

Harsh Chandak
KIT's College of Engineering (Autonomous)
Kolhapur, Maharashtra, India
harshchandak124@gmail.com

Srushtirani Patil
KIT's College of Engineering (Autonomous)
Kolhapur, Maharashtra, India
srushtiranipatil2008@gmail.com

Janhavi Parkar
KIT's College of Engineering (Autonomous)
Kolhapur, Maharashtra, India
janhaviparkar2005@gmail.com

Deepanshu Jad
KIT's College of Engineering (Autonomous)
Kolhapur, Maharashtra, India
deepanshujad@gmail.com

***Abstract*—This research explores the application of machine learning techniques in customer analytics, specifically focusing on customer behavior prediction and segmentation. With the rise of e-commerce and the digitization of retail, businesses are accumulating vast amounts of customer data. Leveraging this data is critical for improving customer experience, enhancing targeted marketing, and increasing customer retention. The study uses a quantitative approach, employing exploratory data analysis (EDA) and machine learning algorithms, such as K-Means clustering, on transactional data obtained from a public dataset of 56,251 records across 17 columns. The analysis identifies trends in purchasing behavior and enables segmentation for more personalized marketing strategies. Key findings demonstrate the value of clustering algorithms in segmenting customers, facilitating better business decisions, and enhancing customer satisfaction. The research also highlights the potential of interactive visualizations in providing insights and driving customer engagement strategies. Limitations of the study, such as missing demographic information and the temporal scope of data, are discussed, along with ethical considerations for data privacy.**

***Index Terms*—Customer Analytics, Data Analysis .**

I. Introduction

Within the context of business today, businesses use more and more data in order to make decisions that improve customer experience and increase sales. The growth of e-commerce and digitization of conventional retail has created volumes of customer data, allowing retailers to learn more about consumers. One of the most important areas of data science is retail customer analytics, which enables businesses to look beyond the surface of transactional records or loyalty programs or other sources and see patterns in all interactions done online by customers. It enables businesses to undertake targeted marketing campaigns, make better inventory forecasting, and improve customer loyalty through analysis of identified patterns in data. Nevertheless, even if the ready available data is increasing on a daily basis, many retailers have difficulties employing customer analytics to the fullest because of the intricacies of the data devolved to volumes and the advanced level of sophistication in predictive techniques modeled. In this research paper, the application of machine learning capabilities in retail customer analytics will be analyzed in order to inform prediction of customer behaviour and effective customer segmentation. The aim of this study is to use data gathered in a retail setting to indicate what business tactics and resolutions along with their implementation will lead to enhanced consumer satisfaction through the use of predictive models.

II. Related Work

In the realm of customer segmentation and analytics, various studies have explored clustering techniques like K-Means alongside models such as RFM (Recency, Frequency, Monetary) to enhance business strategies and boost customer retention. For example,in paper [1] authors utilized the K-Means clustering algorithm in conjunction with the RFM model on a UK e-commerce dataset, categorizing customers into segments like high-value, intermediate, and at-risk, and confirming the results with a silhouette index score of 0.442, which indicates strong clustering performance. Similarly,in paper [2] the employed RFM analysis and K-Means to examine real-time transactional datasets, assessing customer buying patterns across different regions and identifying optimal clusters through the silhouette coefficient. Furthermore, paper [3] investigated factors affecting customer satisfaction, demonstrating how secure payment options and consumer trust can enhance retention and drive business growth, which are crucial elements in customer segmentation strategies.Paper [4] highlighted the significance of hybrid clustering methods in CRM strategies, pointing out that integrating algorithms like K-Means with other techniques can yield more precise insights for analyzing consumer behavior and executing targeted marketing.

These studies lay the groundwork for understanding and applying clustering in customer segmentation, ultimately leading to improved retention and profitability.

TABLE I: Literature Review

Year	Author	Title	Methodology
2021	Rahul Shirole, Laxmiputra Salokhe, Saraswati Jadhav	Customer Segmentation using RFM Model and K-Means Clustering [1]	Exploratory Data Analysis (EDA) helps identify unique customers, order percentages, and data inconsistencies, such as missing or null values. After preprocessing, RFM analysis is performed by creating Recency (days since last purchase), Frequency (time between purchases), and Monetary (amount spent) variables. K-Means algorithm using Euclidean distance is applied twice to segment customers: first based on recent transactions and second on frequent transactions. The resulting clusters are evaluated using the Silhouette score, and comparisons are made for K = 3 and K = 5. Finally, the clusters are analyzed based on sales recency, frequency, and amount to identify high-value customer groups.
2021	Nazmun Nessa Moon, Iftakhar Mohammad Talha, Imrus Salehin	An advanced intelligence system in customer online shopping behavior and satisfaction analysis [2]	The quality of the product, the price of the product compared to the local market, the policy of return, timely delivery of the product are also essential elements of online shopping.By analyzing all these factors,by using Apiorir algorithm Naive Bayes they have tried to find the customer behavior and satisfaction with online shopping.
2022	P. Anitha , Malini M. Patil	RFM model for customer purchase behavior using K-Means algorithm [3]	a)Exploratory analysis and data processing. b)Execution of RFM analysis c)K-means algorithm is applied using Euclidean distance metric to partition the customers for RFM values. d)calculation of silhouette score.
2020	Patel Monil, Patel Darshan, Rana Jecky, Chauhan Vimarsh, Prof. B. R. Bhatt	Customer Segmentation using Machine Learnin [4]	This paper is about for Predicting the future consumption trend of customers in the way of segmentation of customer information based on geographic,demographic,psychographic and consumption behavior.And Profit market planning of enterprises, so as to achieve the goal of reasonable allocation of service resources and the most profitable design of customer marketing programs. For that they used. Customer Relationship Management(CRM),Customer Segmentation,Clustering : K-means clustering, hierarchical clustering, density based clustering,affinity propagation algorithm.
2023	Abdalwali Lutfi, Mahmaod Alrawad, Adi Alsyouf , Mohammed Amin Almaiah ,Ahmad Al-Khasawneh , Akif Lutfi Al-Khasawneh ,Ahmad Farhan Alshira,Malek Hamed Alshirah , Mohamed Saad , Nahla Ibrahim	Drivers and impact of big data analytic adoption in the retail industry: A quantitative investigation applying structural equation modeling [5]	This study used a quantitative questionnaire with items adapted from previous research and rated on a five-point Likert scale. After reviewing by academicians and industry experts, the questionnaire was translated into Arabic and pre-tested for reliability, with Cronbach's alpha exceeding 0.70. The research hypotheses were formulated from the literature, with BDA adoption and exogenous factors based on the TOE model. The target population was 500 SME retail firms in Jordan, with 137 responses, 132 of which were usable (26.4% response rate). Non-response bias was tested and found insignificant.

2021	V.Vijilesh, A.Harini, M.Hari Dharshini, R.Priyadharshini	CUSTOMER SEGMENTATION USING MACHINE LEARNING [6]	This paper shows how innovation and customer understanding are vital for successful e-commerce. With the overwhelming variety of products available, businesses can use customer segmentation to reduce confusion and improve targeting. By grouping customers into High, Medium, and Low segments based on purchasing behavior, machine learning, particularly the k-means clustering algorithm, helps uncover patterns in data. This approach allows B2C companies to enhance their decision-making, develop customized products and services, and better compete in the market.
2022	Desi Adrianti Awaliyah1,Budi Prasetiyo, Rini Muzayanah, Apri Dwi Lestari	Optimizing Customer Segmentation in Online Retail Transactions through the Implementation of the K-Means Clustering Algorithm [7]	This paper shows that using the Recency, Frequency, and Monetary (RFM) approach, combined with the K-means algorithm, optimizes customer segmentation, enabling companies to better understand and meet customer needs. The study employs the elbow method to determine that three clusters are optimal for segmenting online retail customers. The analysis highlights that factors like quantity and unit price significantly impact customer behavior. Additionally, the research introduces new variables to improve transaction behavior insights. The paper concludes that effective customer segmentation aids in business decision-making, such as offering rewards, and suggests future research to compare K-means with other algorithms for better performance.
2022	VAMSI KATRAGADDA	Dynamic Customer Segmentation: Using Machine Learning to Identify and Address Diverse Customer Needs in Real-Time [8]	This paper shows that machine learning can significantly improve customer segmentation by dynamically adapting to changing behaviors and preferences. The study compares algorithms like k-means and decision trees, finding that machine learning models outperform traditional segmentation methods in accuracy and responsiveness. Using real-time data, businesses can personalize marketing strategies and enhance customer interactions, leading to higher satisfaction and profitability. The research highlights the importance of dynamic segmentation in modern marketing and suggests future exploration into predictive analytics, unstructured data integration, and ethical considerations in AI-driven marketing.
2021	Prof. Nikhil Patankar, Soham Dixit, Akshay Bhamare, Ashutosh Darpel and Ritik Raina	Customer Segmentation Using Machine Learning [9]	The goal is to use behavioral traits (income and expenditure) to categorize consumers for more focused advertising. Method: For effective consumer segmentation, the K-means clustering method is used. Method: collection and cleansing of data. Extraction of features (annual income, spending score). The Elbow technique is used for clustering and hyperparameter tweaking. Result: Developed clusters allow for customized marketing tactics. Cons: Higher marketing expenses and maybe lower productivity because of tiny sector sizes.

2021	Vasant Dhar	Data Science Prediction [10]	This article discusses how big data and machine learning are changing fields such as medicine, social sciences, and politics. It underlines predictive modeling, explains contrasts between causal vs. predictive accuracy, discusses the shift from hypothesis-driven to data-intensive research, and makes opportunities such as large-scale randomized experiments.
2022	Chenguang Wang	Efficient customer segmentation in digital marketing using deep learning with swarm intelligence approach [11]	The concept under discussion in this research paper is an AI-based model for customer segmentation in digital marketing. It considers using deep learning and swarm intelligence to more efficiently group customers, considering their purchasing patterns. This model incorporates feature selection, clustering, and classification procedures to analyze data on the customers and further enhance business growth. A comparison of the suggested model with the existing models is done, demonstrating high performance by this proposed model for customer segmentation. In addition, it offers a comprehensive review of related literature and motivation into the need for proper customer segmentation strategies. Conclusion The paper concludes by outlining the proposed methodology and potential impact on digital marketing.
2022	Kamil Matuszela ´nski and Katarzyna Kopczewska	Customer Churn in Retail E-Commerce Business: Spatial and Machine Learning Approach [12]	The forthcoming research predicts retail e-commerce customer churn using order, reviews, and demographic information. It identifies that topic modeling, spatial clustering, and machine learning with the use of XGBoost outperformed traditional methods. The key churn indicators proved to be initial order value, product category, and demographics, where review scores and location type matter less. Based on findings, the study highlights the requirement of behavioral data in retention decisions and further researches through the integration of machine learning models into geodemographic analysis for development of effective CRM strategies.
2014	Hasan Ziafat , Majid Shakeri	Using Data Mining Techniques in Customer Segmentation [13]	This research aims at customer segmentation using data mining techniques to categorize customers into separate groups for targeted marketing strategies. Major techniques include clustering algorithms which identify customer segments, analyzing behaviors, and value-based segmentation. In the segmentation process, understanding context, preparing data, modeling, and clustering the customers, evaluating the result, and implementing strategies. Recommendations include using supervised models, cleaning data, applying multiple models for accuracy, and careful labeling of customer segments. These are to help tailor the marketing efforts and improve the management of customer relationships.

III. Methodologies

This research aims to explore customer shopping behaviors through a comprehensive analysis of transactional data using Exploratory Data Analysis (EDA) and interactive visualizations. The quantitative approach allows for the identification of trends, customer segmentation, and potential marketing strategies based on observed behaviors.

A. Research Design

The study follows a quantitative research design, utilizing secondary data from a large e-commerce platform. Exploratory data analysis (EDA) techniques were employed to uncover insights into customer purchasing patterns, while interactive visualizations were created using the Plotly library in Python. This allows for dynamic exploration of the results.

B. Data Collection

Data collection refers to the process of gathering and measuring data from various sources for the purpose of analyzing, interpreting, and using it to make decisions. Data collection is a systematically obtained process of data related to a particular purpose, such as research analysis, business studies, or solving a problem by being accurate and reliable. Methods of collecting data include surveys, interviews, observations, and the use of existing databases or digital tracking tools.

C. Data Preprocessing

Data Cleaning: Data cleaning was performed to ensure accuracy and consistency in the dataset. This involved:

Handling Missing Values: Missing values in the Age field were imputed with the mean age to retain as many records as possible. [7]

Duplicate Records: Duplicates were identified and removed to avoid skewing the results.But no duplicates were found. [7]

Date Parsing: The order dates field analyzed over different time frames such as months or years.

D. Exploratory Data Analysis (EDA)

EDA was conducted to explore and understand the key characteristics of the data. This included:

Univariate Analysis: Descriptive statistics were calculated to understand the distribution of individual variables

E. Visualization with Power BI

The Power BI was used to create interactive visualizations that allowed for a deeper exploration of the dataset:

Bar Charts: Visualized the total spending per product category, helping identify which product categories contributed the most to revenue.

Histograms: Displayed the distribution of customer spending and shopping malls, highlighting spending patterns across different demographics.

Pie Charts: Understanding the distribution of payment methods through a pie chart or data analysis offers several benefits, particularly for businesses, marketers, and financial analysts.

F. Data Analysis Methods

Statistical Analysis: Descriptive statistics (mean, median, standard deviation) were used to summarize key variables. A Pearson correlation test was also used to measure the strength of relationships between variables, such as age and spending behavior.

G. Software and Tools

The following software and tools were used in the analysis:

Python: The primary programming language used for data manipulation, analysis, and visualization.

Pandas: For data cleaning, preprocessing, and analysis.

Power BI: For creating interactive visualizations.

Scikit-learn: For clustering and machine learning tasks.

Jupyter Notebook: For organizing the code, analysis, and visualizations in an interactive environment.

Power BI: for creating interactive dashboards and reports, visualizing data, analyzing trends, integrating data from multiple sources, and supporting real-time business decision-making.

IV. Results and Discussions

A. Data Collection

This paper uses an online retail transaction dataset, which in itself contains 56,251 entries with 17 columns. The dataset is the base from which the analysis and research are drawn from. The variables to be the utilization of this study are outlined and detailed in figure 1

	url	address	name	online_order	book_table	rate	votes	phone	
0	https://www.zomato.com/bangalore/jalsa-banasha...	942, 21st Main Road, 2nd Stage, Banashankari, ...	Jalsa	Yes	Yes	4.1/5	775	080 42297555\r\n+91 9743772233	Ban
1	https://www.zomato.com/bangalore/spice-elephan...	2nd Floor, 80 Feet Road, Near Big Bazaar, 6th ...	Spice Elephant	Yes	No	4.1/5	787	080 41714161	Ban
2	https://www.zomato.com/SanchurroBangalore?cont...	1112, Next to KIMS Medical College, 17th Cross...	San Churro Cafe	Yes	No	3.8/5	918	+91 9663487993	Ban
3	https://www.zomato.com/bangalore/addhuri-udupi...	1st Floor, Annakuteera, 3rd Stage, Banashankar...	Addhuri Udupi Bhojana	No	No	3.7/5	88	+91 9620009302	Ban
4	https://www.zomato.com/bangalore/grand-village...	10, 3rd Floor, Lakshmi Associates, Gandhi Baza...	Grand Village	No	No	3.8/5	166	+91 8026612447\r\n+91 9901210005	Basa

figure 1

B. Data Preprocessing

Data Transformation: Data from the pool is processed in preparation for further analysis at the data analysis and stage. At this stage, the columns URL, address, phone, menu items, dish liked, review list are removed to avoid unnecessary processing. Type of dataset before and after data transformation As shown in Figure 1 and Figure 2.

	name	online_order	book_table	rate	votes	location	rest_type	cuisines	approx_cost(for two people)	listed_in(type)	listed_in(city)
0	Jalsa	Yes	Yes	4.1/5	775	Banashankari	Casual Dining	North Indian, Mughlai, Chinese	800	Buffet	Banashankari
1	Spice Elephant	Yes	No	4.1/5	787	Banashankari	Casual Dining	Chinese, North Indian, Thai	800	Buffet	Banashankari
2	San Churro Cafe	Yes	No	3.8/5	918	Banashankari	Cafe, Casual Dining	Cafe, Mexican, Italian	800	Buffet	Banashankari
3	Addhuri Udupi Bhojana	No	No	3.7/5	88	Banashankari	Quick Bites	South Indian, North Indian	300	Buffet	Banashankari
4	Grand Village	No	No	3.8/5	166	Basavanagudi	Casual Dining	North Indian, Rajasthani	600	Buffet	Banashankari

Figure 2

Handling missing value: The following step in data exploration would be to check for missing data that may be present in the dataset, also known as empty values, which are absent or undefined values in the dataset. In this study, missing values are addressed by removing rows that have a missing value or by adding the mean value of the correlated values.

After all the data preprocessing and analyzing the dataset to work on is created, shown in Figure 3.

	name	online_order	book_table	rate	votes	location	rest_type	cuisines	Cost2plates	Type
0	Jalsa	Yes	Yes	4.1	775	Banashankari	Casual Dining	North Indian, Mughlai, Chinese	800.0	Buffet
1	Spice Elephant	Yes	No	4.1	787	Banashankari	Casual Dining	others	800.0	Buffet
2	San Churro Cafe	Yes	No	3.8	918	Banashankari	others	others	800.0	Buffet
3	Addhuri Udupi Bhojana	No	No	3.7	88	Banashankari	Quick Bites	South Indian, North Indian	300.0	Buffet
4	Grand Village	No	No	3.8	166	Basavanagudi	Casual Dining	others	600.0	Buffet

Figure 3

Data Visualization:

Visualizing data with Seaborn and Matplotlib will allow insight to be extracted from restaurant data with a combinationof different plots:

Bar Chart: Online Orders vs Count Online orders are presented across categories. By analyzing it, demand trends for online ordering can be seen, which helps in optimizing online channels.

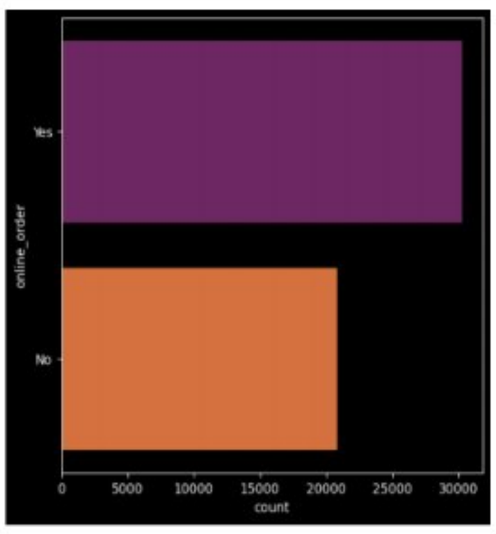

Figure 4: Online Orders vs Count Online

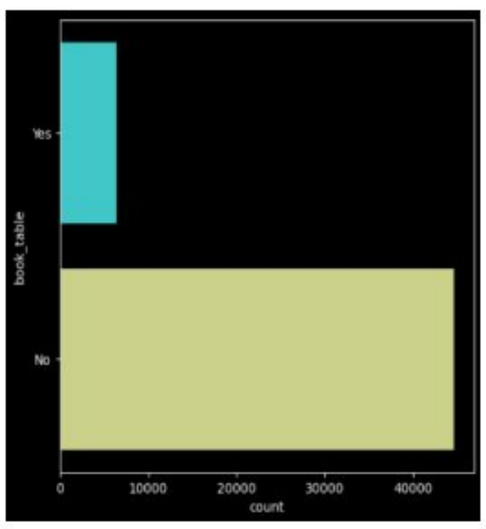

Figure 5: Book Table vs Count of table bookings

Bar Chart: Book Table vs Count of table bookings is presented, and it helps in identifying peak booking times to enhance reservation management.

Box Plot: Rating vs Online Order It provides the distributionof customer ratings on online orders for analyzing customer satisfaction related with online ordering.

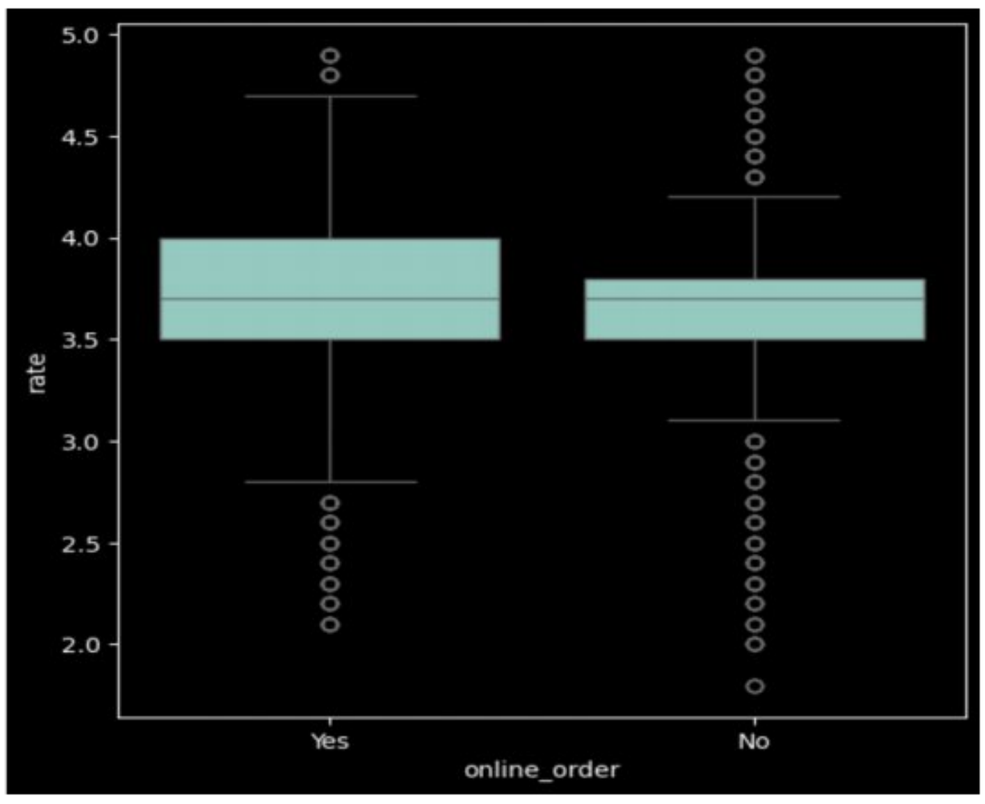

Figure 6

Box Plot: Rating vs Table The plot shows how and whether or not ratings are associated with table bookings, indicating customer satisfaction related to in-restaurant experiences.

Location Plot The map plots restaurant locations, allows checking about the distribution, customer reach, and regional trends and supports strategic decision-making.

These plots enable restaurants to make better decisions, including operations improvement.

A dashboard in Power BI with locations, cuisine type, word cloud, restaurant category, and number of votes and ratings clearly delivers the restaurant data with interactivity.

Locations: Maps expose distribution and activity of restaurants across locations.

Cuisine Type: Available in the form of charts that allow filtering for in-depth information.

Cuisines Word Cloud: Displays the frequency of cuisines visually, emphasizing popular ones.

Restaurant Types: This categorizes the restaurants into types buffets, deserts, etc.

Votes and Ratings: Tracks customer feedback and preferences, helping identify top-performing restaurants.

V. Conclusion

Visualization is an important customer data analytics tool by which organizations easily discover trends and patterns while making timely decisions. Interactive abilities in Power BI present a scalable, cost-effective, and user-friendly platform for big data and complex data handling. Its ability to integrate with multiple sources of data and to provide real-time updates places it in a very desirable position in highly dynamic business environments. As this research indicates, visualization has the potential

Clean data: Make raw data meaningful and user-friendly through visuals. Inform decisions: Offer actionable insights with crisp metrics and trends. Improve engagement: Visualize customer behavior and segments for tailored strategies. Empower teams: Foster cooperation through shared dashboards. Drive predictions: Pinpoint opportunities and risks

from historical data trends. Power BI shows how advanced visualizations transform raw data into strategic insights and enable organizations to stay ahead in a data-driven world.

REFERENCES

[1] Rahul Shirole, Laxmiputra Salokhe, Saraswati Jadhav, "Customer Segmentation using RFM Model and K-Means Clustering," 2021 International Journal of Scientific Research in Science and Technology, 2021, doi : https://doi.org/10.32628/IJSRST2183118.

[2] Nazmun Nessa Moon, Iftakhar Mohammad Talha, Imrus Salehin, "An advanced intelligence system in customer online shopping behavior and satisfaction analysis," 2021 Current Research in Behavioral Sciences,2021, doi:https://doi.org/10.1016/j.crbeha.2021.100051.

[3] P. Anitha , Malini M. Patil, "RFM model for customer purchase behavior using K-Means algorithm," 2022Journal of King Saud University -Computer and Information Sciences, 2022, doi : https://doi.org/10.1016/j.jksuci.2019.12.011.

[4] Patel Monil, Patel Darshan, Rana Jecky, Chauhan Vimarsh, Prof. B. R. Bhatt, "Customer Segmentation using Machine Learnin," 2020 International Journal for Research in Applied Science and Engineering Technology (IJRASET), 2020, doi: http://doi.org/10.22214/ijraset.2020.6344.

[5] Abdalwali Lutfi,Mahmaod Alrawad,Adi Alsyouf, Mohammed AminAlmaiah ,AhmadAl-Khasawneh , Akif Lutfi Al-Khasawneh,Ahmad FarhanAl- shira,Malek Hamed Alshirah , Mohamed Saad , Nahla Ibrahim, "Drivers and impact of big data analytic adoption in the retail in- dustry: A quantitative investigation applying structural equation modeling ," 2023 Journal of Retailing and Consumer Services, 2023, doi:https://doi.org/10.1016/j.jretconser.2022.103129.

[6] V.Vijilesh, A.Harini, M.Hari Dharshini, R.Priyadharshini, "CUSTOMER SEGMENTATION USING MACHINE LEARNING," 2021 International Research Journal of Engineering and Technology (IRJET), 2021,Volume: 08 Issue: 05.

[7] Desi Adrianti Awaliyah, Budi Prasetiyo, Rini Muzayanah, Apri Dwi Lestari, "Optimizing Customer Segmentation in Online Retail Transactions through the Implementation of the K-Means Clustering Algorithm ," 2024 Scientific Journal of Informatics Vol. 11, No. 2,May 2024 doi:10.15294/sji.v11i2.6137.

[8] VAMSI KATRAGADDA, "Dynamic Customer Segmentation: Using Machine Learning to Identify and Address Diverse Customer Needs in Real-Time ," 2022 ICONIC RESEARCH AND ENGINEERING JOURNALS , 2022, IRE 1703349

[9] Prof. Nikhil Patankar,Soham Dixit,Akshay Bhamare, Ashutosh Darpel and Ritik Raina, "Customer Segmentation Using Machine Learning " , 2021, doi:10.3233/APC210200

[10] Vasant Dhar , "Data Science Prediction " , 2021, doi:10.1145/25004999

[11] Chenguang Wang , "Efficient customer segmentation in digital marketing using deep learning with swarm intelligence approach" , 2022, https://doi.org/10.1016/j.ipm.2022.103085

[12] Kamil Matuszela´nski and Katarzyna Kopczewska , "Customer Churn in Re- tail E-Commerce Business: Spatial and Machine Learning Approach" , 2022, . https://doi.org/ 10.3390/jtaer17010009

[13] Hasan Ziafat , Majid Shakeri, "Using Data Mining Techniques in Customer Segmentation " , 2014, Hasan Ziafat Int. Journal of Engineering Research and Applications ISSN : 2248-9622, Vol. 4, Issue 9(Version 3), September 2014, pp.70-79