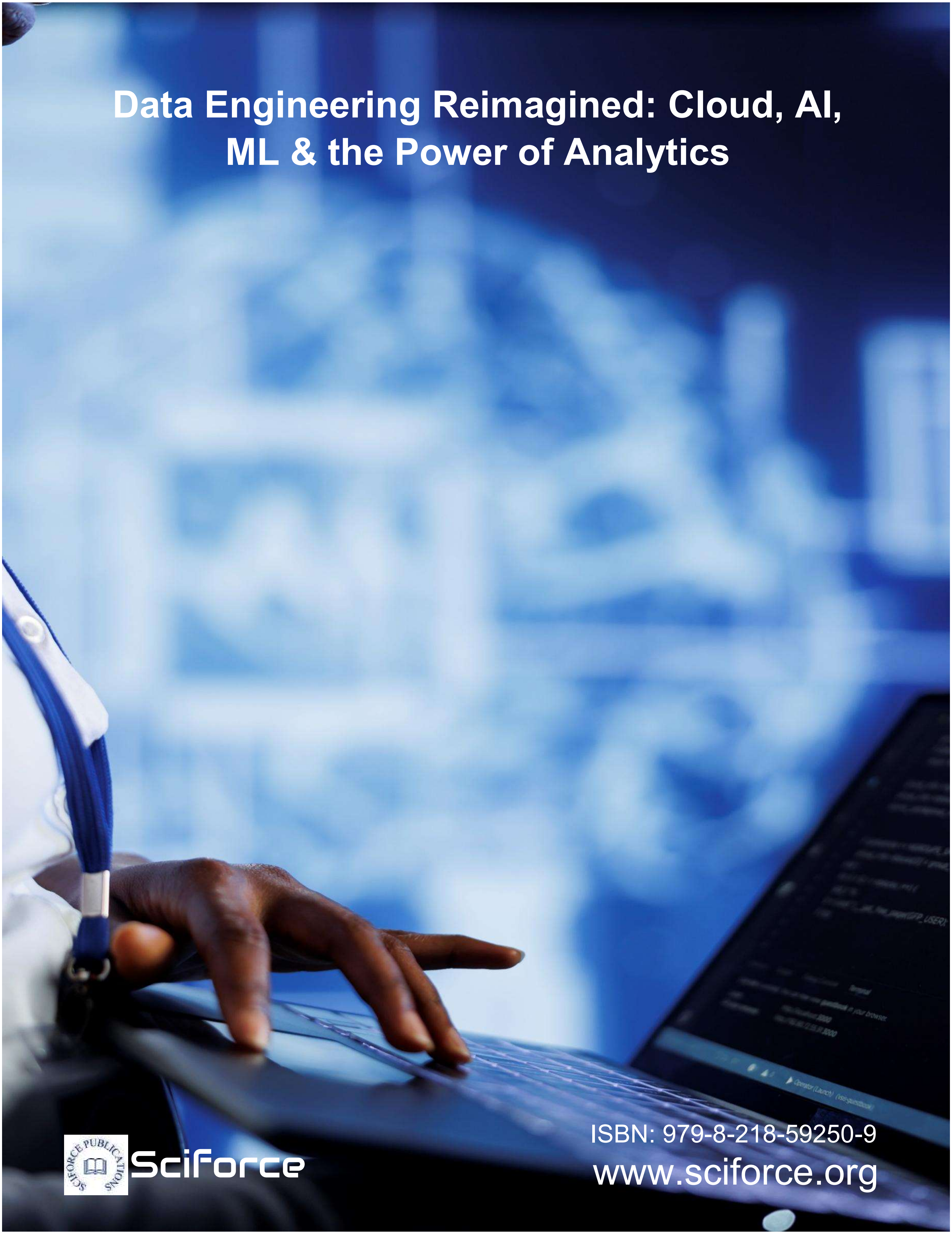
Data Engineering Reimagined: Cloud, AI, ML & the Power of Analytics



ISBN: 979-8-218-59250-9
SciForce Publications
SciForce
www.sciforce.org

Sciforce Publications
www.sciforce.org

Customer Personality Analysis Using K-Means Clustering and Ada Boost Regression: A Machine Learning Approach to Market Segmentation

Authors & Editors

Divya Soundarapandian

Correspondence

Software Engineering Manager, The Home Depot., United State

Published By
Sciforce Publications
April 20, 2025

Sciforce Publications
www.sciforce.org

Customer Personality Analysis Using K-Means Clustering and Ada Boost Regression: A Machine Learning Approach to Market Segmentation

This study explores customer personality analysis using unsupervised machine learning techniques, which can help improve business planning and marketing effectiveness. In today's competitive business landscape, understanding customer behavior and preferences is essential for organizational success. Traditional market segmentation based on basic demographics provides limited insights into consumer behavior, necessitating a more sophisticated approach. This research uses the k-means clustering algorithm to analyze customer characteristics, behaviors, and preferences to segment customers into well-defined groups. The study examines key metrics including average spend score, annual revenue, total purchases, and customer loyalty score to identify distinct customer segments with distinct personality traits. By integrating data from multiple sources such as online activity, purchase histories, and social media interactions, the research uncovers the psychological and emotional drivers behind purchasing decisions. Advanced visualizations including histograms, bar charts, and doughnut charts help provide a deeper understanding of customer profiles. The analysis identifies five basic personality traits: extraversion, openness, conscientiousness, agreeableness, and neuroticism. This approach helps businesses shift from generic marketing strategies to precisely targeted campaigns that resonate with specific customer segments. The study proposes a personality-based product recommendation system that aims to improve customer engagement, satisfaction, retention, and profitability, while addressing challenges such as customer confusion and fraudulent profiles.

Keywords: Customer Persona Analysis, Unsupervised Learning, K-means clustering, Customer Segmentation. Machine Learning, Customer Loyalty Scoring, Purchasing Behavior, Predictive Analytics, Target Marketing, Ada Boost Regression.

INTRODUCTION

A fundamental component of modern business planning is customer persona analysis, which helps companies gain a deeper understanding of their customers and improve service. For this task, unsupervised learning is a useful machine learning method that allows them to examine large amounts of customer data to find common themes and unique characteristics across different consumer segments [1]. Clustering techniques help businesses group consumers with similar characteristics into distinct customer segments, each defined by unique personality traits. This segmentation allows them to precisely target marketing and sales efforts towards each specific group. Furthermore, the ability of unsupervised learning algorithms to analyze vast datasets makes them essential for uncovering deep and actionable customer insights [2].

This paper describes customer persona analysis, which involves a detailed examination of a company's most valuable customers by studying their characteristics, behaviors, and preferences. The purpose of this method is to improve business performance by understanding the complexities of the customer base, thereby allowing for product customization that directly addresses the specific needs and concerns of unique customer segments [3]. The study uses advanced visualizations such as histograms, bar charts, doughnut charts, and table charts to gain a deeper understanding of customer profiles. The aim is to facilitate a strategic shift from a generic, homogeneous model to a precisely targeted approach, which improves resource allocation and helps design more effective marketing campaigns that resonate with specific customer segments [4].

In the contemporary business landscape characterized by fierce competition, the ability to understand customer behavior and preferences is fundamental to any business that wants to thrive. This study addresses this imperative by segmenting consumers into well-defined groups using a dual approach of customer persona analysis and machine learning. Using the k-means clustering algorithm, the research aims to identify sophisticated patterns and correlations that link customer profiles to their purchasing habits [5]. By analyzing customer interactions, demographic information, and psychological data, key personality traits, including extraversion, openness, conscientiousness, agreeableness, and neuroticism, are identified. The purpose of this analysis is to provide actionable insights that enable targeted marketing, increase customer satisfaction, and ultimately drive greater profitability [6].

Traditional market segmentation based on basic demographics such as age and gender provides a limited view of consumer behavior. To address this shortcoming, customer persona analysis examines the psychological and emotional drivers behind purchasing decisions. By integrating data from a variety of sources such as online activity, purchase histories, and social media, companies can uncover the underlying motivations and thought processes of their customers. Powered b machine learning and predictive analytics, this cutting-edge methodology helps companies predict future consumer actions, design more powerful strategies, and fundamentally reshape modern marketing practices [7].

We propose a product recommendation system enhanced by a customer personality assessment tool. By integrating personality-based insights, the system provides recommendations with the best fit for each individual. This approach results in a significant difference in the products recommended to customers who, under conventional systems, received similar recommendations [8]. Customer personalization analytics is emerging as a key strategy that enables businesses to segment vast customer datasets and, as a result, personalize their offerings to unique consumer types. This targeted approach directly improves customer engagement and fosters greater brand loyalty. In today's digital world, accurately characterizing customer personality types is a major challenge for technology companies.

This challenge is particularly important because these companies often suffer significant financial losses due to customer churn. Mitigating this churn and building lasting customer loyalty, especially with the added problem of fraudulent profiles, relies heavily on early detection of these personality traits. As a result, a significant body of prior research has been devoted to analyzing the causes of customer churn and developing effective strategies for managing it [9]. A brand’s personality is one of the fundamental elements of its identity, and it is crucial for effective communication. Without a unique personality, a brand has no way of expressing itself in a memorable way to consumers. A coherent and consistent brand image ensures that it remains in the mind of the customer. This image is a concept created through rational and emotional feelings. While companies use culture and strategic positioning to introduce their brand, it is ultimately the customer’s personal interpretation of these elements that defines the brand image [10].

MATERIALS AND METHOD

Average Spending Score: The average spend score is a metric used to assess the purchasing behavior of customers over a period of time. It represents the average value of individual spend scores, which are typically assigned based on factors such as purchase frequency, transaction value, and loyalty. This score helps businesses identify high- and low-value customers, segment markets, and design targeted marketing strategies. A higher average spend score indicates greater customer engagement and profitable potential for the company.

Annual Income (k$): Annual income (k$) refers to the total money an individual earns in a year, expressed in thousands of dollars. It includes all sources of income, such as salary, bonuses, investments, and other earnings before taxes and deductions. In data analytics and marketing research, annual income is an important demographic variable used to understand the purchasing power, financial capability, and spending behavior of customers. It helps businesses segment markets and design products or services according to income levels.

Total Purchases: Total purchases refer to the total number or monetary value of all transactions made by a customer over a given period of time. It reflects a customer's overall purchasing activity and engagement with a brand or business. This metric helps companies analyze purchasing patterns, identify loyal customers, and assess sales performance. In marketing and customer service, total purchases are used to assess consumer behavior, measure profitability, and develop targeted strategies to increase customer retention and revenue growth.

Customer Loyalty Score: A customer loyalty score is a metric that measures the strength of a customer's relationship with a brand and their commitment to it over time. It is typically derived from factors such as purchase frequency, repeat transactions, engagement level, and customer satisfaction. A higher loyalty score indicates that customers are more likely to continue to purchase, recommend the brand, and resist switching to competitors. Businesses use this score to assess retention performance, identify loyal customers, and develop strategies to improve long-term customer relationships and profitability.

Instructions for machine learning

Ada Boost Regression: Ada boost (Adaptive Boosting) is an ensemble algorithm that repeatedly combines weak learners to create a high-performance predictive model algorithm iteratively adds classifiers, focusing on correcting errors from previous samples by giving more weight to misclassified cases. In this article, we will explore how to build the Ada boost algorithm from scratch. This diagram illustrates the working principle of Boosting. Initially, the model is trained on the original dataset. In each subsequent iteration, misclassified data points receive higher weights, leading the next model to prioritize them. This cycle continues until all the models converge to produce a robust, more accurate final prediction. Python provides dedicated libraries for implementing Ada Boost. Next, we will explore how to apply Ada Boost to a machine learning task using Python. For demonstration, we create a synthetic dataset to evaluate its performance on this particular problem.

RESULT AND DISCUSSION

This dataset analyses customer profiles on four key metrics: average spend score (0-100), annual income (in thousands of dollars), total number of purchases, and customer loyalty score. The data generally shows a positive correlation; customers with higher incomes (e.g., ~$130k) have higher spend scores and more purchases. Loyalty scores also often increase with these metrics, with many customers reaching a maximum score of 10. However, there are exceptions, with some low-income customers showing moderate loyalty, while some high-spending individuals have low loyalty scores. This suggests that while income and spend are significant drivers, loyalty is also influenced by other factors not captured here, such as customer service satisfaction or brand alignment.

TABLE 1. Descriptive Statistics

	Average Spending Score	Annual Income (k$)	Total Purchases	Customer Loyalty Score
count	100.0000	100.0000	100.0000	100.0000
mean	56.6800	61.1691	36.0200	5.4925
std	27.2815	30.5732	18.3160	2.6783
min	11.0000	6.5825	6.0000	0.0000
25%	31.0000	32.3946	20.7500	3.3449
50%	60.5000	62.3034	38.0000	6.0632
75%	80.2500	83.9573	49.0000	7.2277
max	99.0000	132.7571	76.0000	10.0000

Table 1 shows, based on descriptive statistics, that the customer base is remarkably diverse in behavior and financial capability. The average customer has a moderate spending score and income, but high standard deviations indicate wide variations. For example, although the average loyalty score is 6.06, the minimum is 0, which represents a segment of customers who are completely disengaged. The data is roughly symmetrical for spending and income (mean ≈ median), but the loyalty scores are slightly left-skewed, meaning that there are more customers than average as a group. This highlights a wide spectrum, from low-value, low-loyal individuals to a core group of high-spending, highly loyal customers.

Effect of Process Parameters

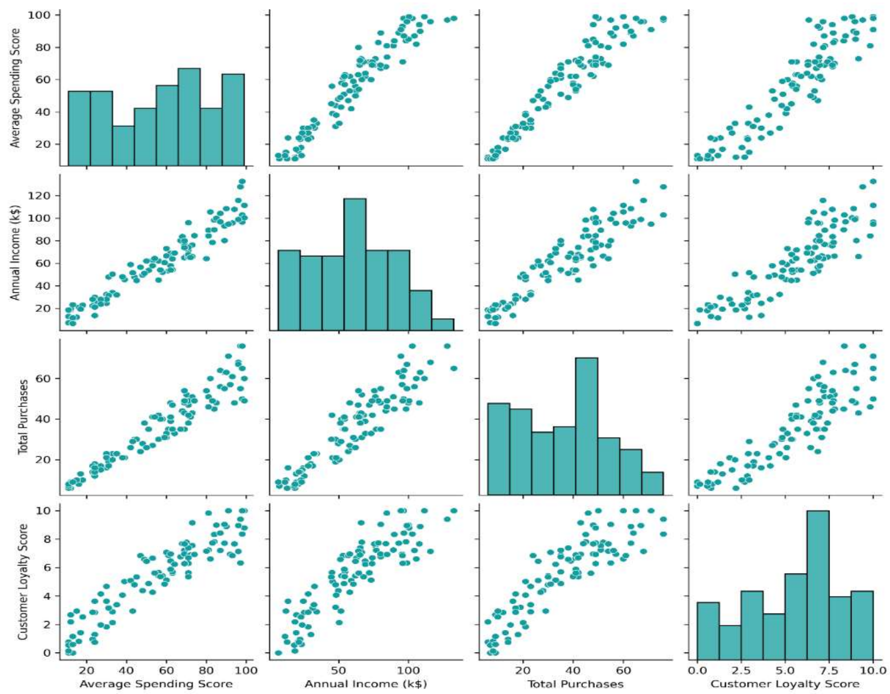


FIGURE 1. Scatter plot of the various Customer Personality Analysis

Figure 1 provides a detailed scatterplot matrix illustrating the relationships between key customer metrics: average spend score, annual revenue, total purchases, and customer loyalty score. The diagonal histograms show the spread of each variable, while the off-diagonal scatterplots reveal positive correlations, particularly between revenue, spend, and purchasing behavior, highlighting patterns in customer persona analysis.

FIGURE 2. Heat map of the relationship between process parameters and outcomes

Figure 2 Description: The heat map illustrates the strong positive correlations between all parameters. Average Spending Score and Annual Income (0.96) exhibit the highest correlation, indicating that customers with higher incomes tend to spend more. Similarly, Total Purchases and Customer Loyalty Score are also positively correlated, indicating that frequent buyers maintain higher loyalty. Overall, the variables are highly interdependent.

Ada Boost Regression (Average Spending Score)

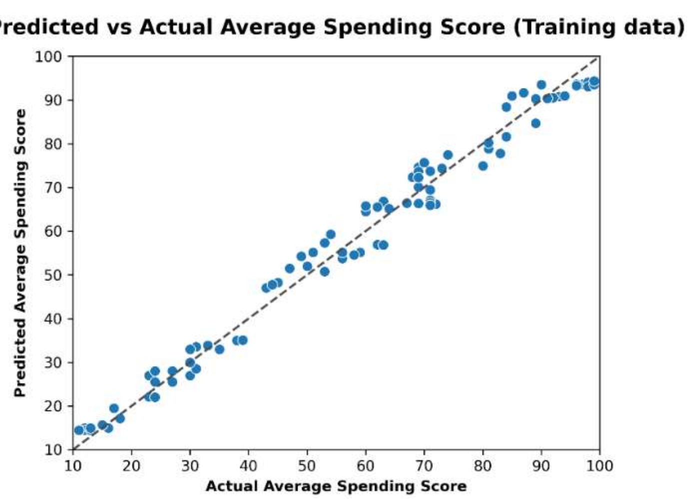


FIGURE 3. Ada Boost Regression on Average Spending Score: training data

Figure 3 Description: The scatter plot shows the performance of the Ada Boost regression model on the training data, comparing the predicted and actual average cost scores. The points closely align with the diagonal line, indicating high prediction accuracy and minimal error. This indicates that the model effectively captures the underlying cost behavior patterns, demonstrating excellent training fit and reliability.

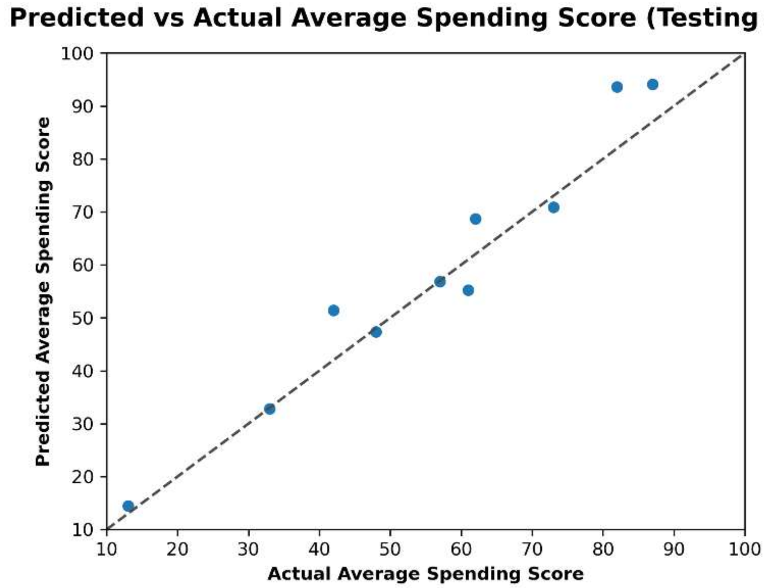


FIGURE 4. Ada Boost Regression on Average Spending Score: testing data

Figure 4 Description: This graph illustrates the performance of the Ada Boost regression model in testing data to predict the average cost score. The predicted values closely match the actual scores on the diagonal line, indicating strong generalization. The minimal deviation from the line indicates that the model maintains high predictive accuracy and reliability on the missing data.

TABLE 2. Performance Metrics of Ada Boost Regression on Average Spending Score (Training Data and Testing Data)

Parameter	Data	Symbol	Model	R2	EVS	MSE	RMSE	MAE	MaxError	MSLE	MedAE
Average Spending Score	Train	ABR	Ada Boost Regression	0.98400	0.98405	12.27694	3.50385	3.15093	6.15385	0.00784	3.30466
	Test	ABR	Ada Boost Regression	0.92219	0.93847	35.99087	5.99924	4.52457	11.65385	0.00937	3.94167

Table 2 shows that the Ada Boost regression model demonstrates excellent performance in predicting customer cost scores, with a strong contrast between its almost perfect fit on the training data and its very good performance on the unseen test data. The high R² value of 0.98 on the training set indicates that the model successfully captured the underlying patterns. Although the test set R² of 0.92 shows a slight expected drop, it confirms the robustness of the model and its ability to generalize to new customers. Key error metrics such as RMSE and MAE are higher for the test set, but at reasonable levels, confirming the model as a reliable and accurate predictive tool.

Ada Boost Regression (Annual Income (k$))

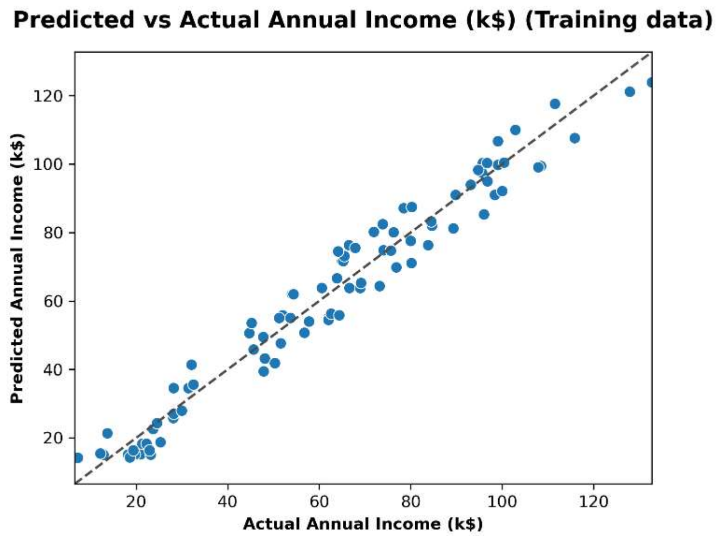


FIGURE 5. Ada BoostRegression on Annual Income (k$): training data

Figure 5 illustrates the performance of the Ada Boost regression model used to predict annual income (in thousands of dollars) using the training dataset. The model combines multiple weak learners, improving prediction accuracy and capturing complex patterns in the data. This visualization highlights how Ada Boost adapts to reduce errors during training.

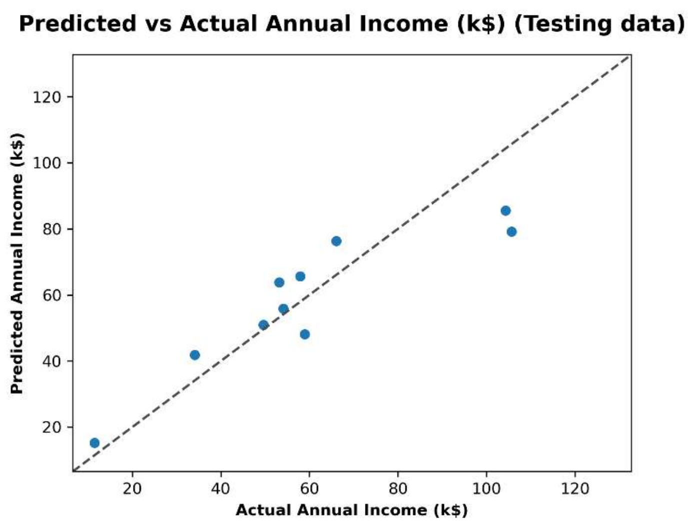


FIGURE 6. Ada Boost Regression on Annual Income (k$)): testing data

Figure 6 presents the Ada Boost regression results on the test dataset for annual income (k$). This plot demonstrates the generalization ability of the model, showing how well it predicts unobserved data. It highlights the consistency and accuracy of Ada Boost in capturing income trends while reducing prediction errors in new models.

TABLE 3. Performance Metrics of Ada Boost Regression on Annual Income (k$) (Training Data and Testing Data)

Parameter	Data	Symbol	Model	R2	EVS	MSE	RMSE	MAE	MaxError	MSLE	MedAE
Annual Income (k$)	Train	ABR	Ada Boost Regression	0.96217	0.96228	35.81795	5.98481	5.25250	10.68633	0.02773	5.79814
	Test	ABR	Ada Boost Regression	0.79181	0.79405	152.01200	12.32932	9.90403	26.39440	0.03370	8.98850

Table 3 shows a significant performance gap between the training and testing phases of the Ada Boost model predicting annual income. On the training data, it achieved a strong R^2 of 0.96, indicating an excellent fit. However, its performance dropped significantly on the unseen test data, with R^2 dropping to 0.79. This significant increase in key error metrics, such as the RMSE increasing from approximately 6 to 12.3 and the MAE increasing from 5.3 to 9.9, indicates that the model is over fitting. It learned too specifically from the training data, including its noise, which reduces its accuracy and generalizability when making predictions on new, real-world customer data.

Ada Boost Regression (Total Purchases)

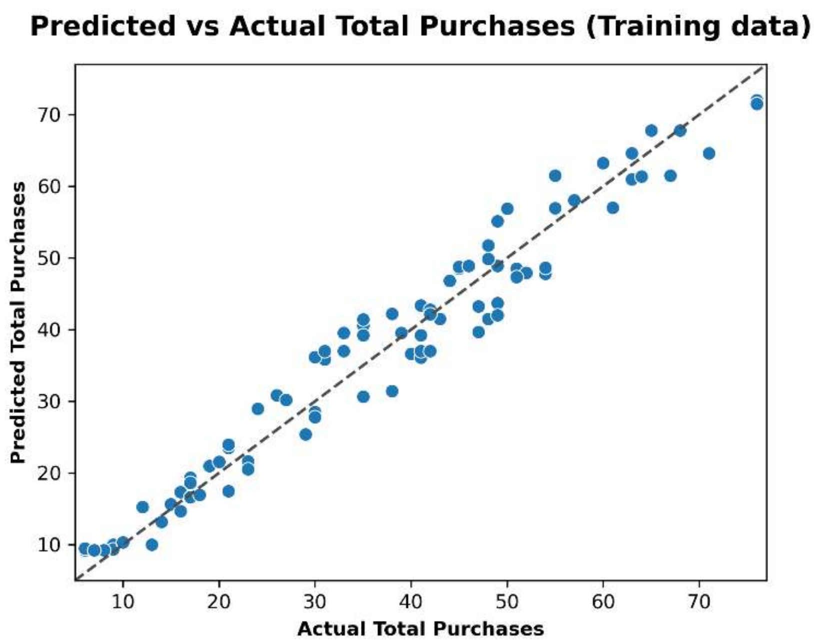


FIGURE 7. Ada Boost Regression on Total Purchases: training data

Figure 7 depicts Ada Boost regression applied to training data for total purchases. The model uses multiple weak learners to accurately capture patterns in customer purchasing behaviour. This visualization demonstrates Ada Boost's effectiveness in fitting the training data, reducing errors, and highlighting relationships between input features and total purchase amounts.

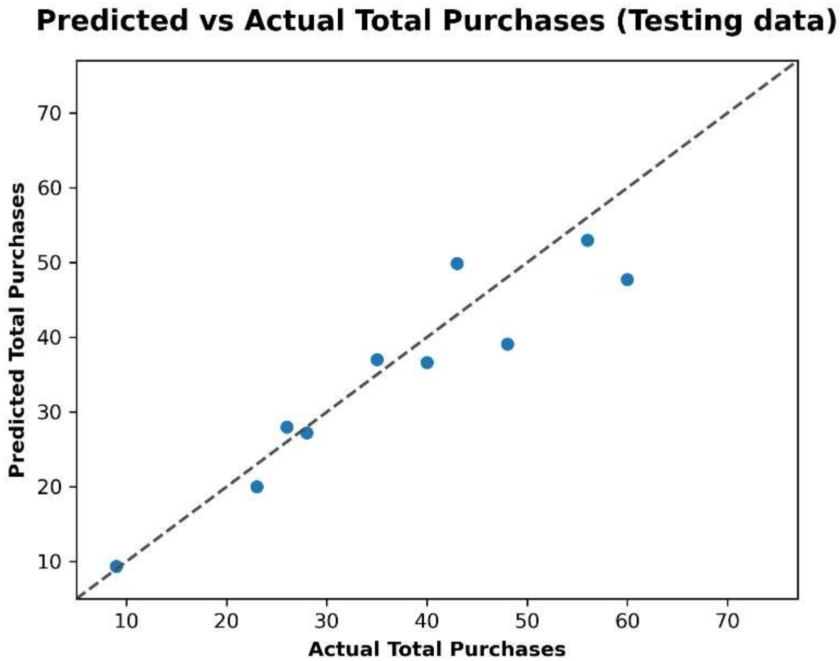


FIGURE 8. Ada Boost Regression on Total Purchases: testing data

Figure 8 shows the Ada Boost regression results on the test dataset for wholesale purchases. This graph evaluates the model's ability to generalize to unobserved data, which illustrates how accurately it predicts customer purchase behavior. It highlights the effectiveness of Ada Boost in maintaining prediction consistency and reducing errors in new models.

TABLE 4. Performance Metrics of Ada Boost Regression on Total Purchases (Training Data and Testing Data)

Parameter	Data	Symbol	Model	R2	EVS	MSE	RMSE	MAE	MaxError	MSLE	MedAE
Total Purchases	Train	ABR	Ada Boost Regression	0.95827	0.95829	14.35662	3.78901	3.21517	7.33333	0.01527	3.10000
	Test	ABR	Ada Boost Regression	0.85925	0.87735	31.55003	5.61694	4.25725	12.27273	0.01499	3.00000

Table 4 Ada Boost model for predicting total purchases demonstrates robust performance, effectively balancing learning and generalization. While the training R² of 0.96 indicates a near-perfect fit, the testing R² of 0.86 confirms that the model retains strong predictive power on new, unseen data. This minimum performance gap indicates minimal over fitting. Key error metrics, such as RMSE (3.79 train vs. 5.62 test) and MAE (3.22 train vs. 4.26 test), show reasonable and expected increases for the testing set. Overall, the model is highly reliable, accurately predicting customer purchase counts with consistent accuracy on both known and new data.

CONCLUSION

Based on the extensive analysis presented in this study, customer personality analysis using unsupervised machine learning techniques, specifically k-means clustering and Ada Boost regression, demonstrates significant potential to transform business marketing strategies and customer engagement approaches. The research successfully identified distinct customer segments characterized by five basic personality traits: extraversion, openness, conscientiousness, agreeableness, and neuroticism, providing actionable insights for targeted marketing campaigns. The Ada Boost regression models showed varying levels of performance across different customer metrics.

With an R^2 of 0.98 on the training data and 0.92 on the test data, the model achieved exceptional accuracy in predicting average spending scores, demonstrating strong generalization capabilities. Similarly, the total purchase prediction model maintained strong performance with a test R^2 of 0.86, indicating minimal over fitting and reliable predictive power. However, the annual income prediction model showed signs of over fitting, with R^2 decreasing from 0.96 in the training data to 0.79 in the test data, highlighting the need for further model refinement and regularization techniques. Descriptive statistical analysis revealed significant heterogeneity in the customer base, with strong positive correlations identified between average spending scores and annual income (0.96) and total purchases and customer loyalty scores. These findings underscore the importance of considering multiple factors beyond the base demographic when developing marketing strategies.

While income and spending patterns are key drivers of customer behavior, the study successfully demonstrated that loyalty is influenced by additional factors such as customer service satisfaction and brand alignment. The proposed personality-based product recommendation system represents a promising approach to improving customer engagement, satisfaction, and retention while addressing contemporary challenges including customer confusion and fraudulent profiles. This research provides the foundation for businesses to shift from generic, uniform marketing approaches to precisely targeted campaigns that resonate with specific customer segments, ultimately resulting in improved resource allocation, increased profitability, and sustainable competitive advantage in today's dynamic business landscape.

REFERENCES

1. Ramadhanti, Alya Rysda, Rafa Syafiq Bastikarana, Andry Alamsyah, and Sri Widiyanesti. "Determining customer relationship management strategy with customer personality analysis using ontology model approach." JournalManagement Indonesia 20, no. 2 (2020): 83-94.
2. Marwade, Anwesh, Nakul Kumar, Shubham Mundada, and Jagannath Aghav. "Augmenting e-commerce product recommendations by analyzing customer personality." In 2017 9th International Conference on Computational Intelligence and Communication Networks (CICN), pp. 174-180. IEEE, 2017.
3. Ahmad, Noman, Mazhar Javed Awan, Haitham Nobanee, Azlan Mohd Zain, Ansar Naseem, and Amena Mahmoud. "Customer personality analysis for churn prediction using hybrid ensemble models and class balancing techniques." IEEE Access 12 (2023): 1865-1879.
4. Ekhlassi, Amir, Majid Hassani Nezhad, Siamak Ashkan Far, and Kamal Rahmani. "The relationship between brand personality and customer personality, gender and income: A case study of the cell phone market in Iran." Journal of Targeting, Measurement and Analysis for Marketing 20, no. 3 (2012): 158-171.
5. Ananda, Lohar Ganesh. "A STUDY OF CUSTOMER PERSONALITY ANALYSIS USING MACHINE LEARNING ALGORITHMS." Statistics and Data Science 108 (2023).
6. Pachchigar, Snehi Nainesh. "Customer personality exploratory data analysis." PhD diss., California State University, Northridge, 2022.
7. Mishra, Vishal, and Sridhar Vaithianathan. "Customer personality and relationship satisfaction: Empirical evidence from Indian banking sector." International Journal of Bank Marketing 33, no. 2 (2015): 122-142.

8. Chandan, Amitabh. "Unveiling Customer Personalities Using Segmentation and Exploratory Data Analysis." Health Leadership and Quality of Life 3, no. 420 (2024): 1.
9. Hee, Ong Choon, and Husna Johari. "A Conceptual Analysis of Personality Traits and Customer-Oriented Behavior in the Health Tourism Hospitals." International Journal of Caring Sciences 7, no. 2 (2014).
10. Peram, S. R. (2025). Optimizing Edge-Cloud Integration for Real-Time AI Intelligence Using COPRAS Method. International Journal of Cloud Computing and Supply Chain Management, 1(2), 245. https://doi.org/10.55124/ijccscm.v1i2.245
11. Chiang, Lan-Lung Luke, and Chin-Sheng Yang. "Does country-of-origin brand personality generate retail customer lifetime value? A Big Data analytics approach." Technological Forecasting and Social Change 130 (2018): 177-187.
12. Chinnasami Sivaji, Vidhya Prasanth, M. Ramachandran, Chandrasekar Raja, "Applications of the MOORA method for Manufacturing Environment", Journal on Electronic and Automation Engineering, 3(4), December 2024, 1-10.
13. Smith, Trevor Alexander. "The role of customer personality in satisfaction, attitude-to-brand and loyalty in mobile services." Spanish Journal of Marketing-ESIC 24, no. 2 (2020): 155-175.
14. Dohan, Din Mohammad, Naheyan Prottush, Sudipta Mondal, Sumaya Binte Zilani Choya, and MD Golam Rabiul Alam. "Customer Personality Analysis Using Machine Learning with Explainable AI." In 2025 13th International Symposium on Digital Forensics and Security (ISDFS), pp. 1-6. IEEE, 2025.
15. Al-Dabbas, Lana, Hassan Al-Tarawneh, and Thamer A. Al-Rawashdeh. "Customer personality segmentation using k-means clustering." In 2023 International Conference on Information Technology (ICIT), pp. 537-543. IEEE, 2023.
16. Menaka, J., & Poulose, A. (2024). The Power of Visual Storytelling: Analyzing Customer Personalities with Tableau. 181–186. https://doi.org/10.1109/icufn61752.2024.10625575
17. Acheme, I. D., & Enoyoze, E. (2024). Customer personality analysis and clustering for targeted marketing. International Journal of Science and Research Archive, 12(1), 3048–3057. https://doi.org/10.30574/ijsra.2024.12.1.1003
18. Kakulavaram, S. R. (2025). “Big Data-Driven Machine Learning: An In-Depth Look at Advantages and Challenges” Journal of Business Intelligence and Data Analytics, vol. 2, no. 2, pp. 1–6. doi: https://dx.doi.org/10.55124/jbid.v2i2.253
19. Mohan, N., Prabhu, M., Nihar, A., & Parveen, E. (2024). The Segmentation Revolution: Customer Personality Analysis Driving Predictive Success. 1, 1775–1778. https://doi.org/10.1109/icaccs60874.2024.10716968
20. Al-Dabbas, L., Al-Tarawneh, H., & Al-Rawashdeh, T. A. (2023). Customer Personality Segmentation Using K-Means Clustering. 537–543. https://doi.org/10.1109/icit58056.2023.10225996
21. Zhou, Mingyuan. "Consumer Personality Analysis: Tailoring Marketing Strategies for Diverse Segments."
22. Peram. S. R (2023). Machine Learning in Wealth Management: Enhancing Investment Strategies through AI. J Comp Sci Appl Inform Technol. 7(1): 1-9
23. Azzadina, Irna, Aulia Nurul Huda, and Corinthias Pamatang Morgana Sianipar. "Understanding relationship between personality types, Marketing-mix factors, and purchasing decisions." Procedia-Social and Behavioral Sciences 65 (2012): 352-357.
24. Kakulavaram. S. R. (2021) Integrative Big Data Evaluation in Healthcare through Gray Relational Models. SOJ Mater Sci Eng 8(2): 1-9. DOI: 10.15226/2473-3032/8/2/00185
25. Castillo, Javier. "The relationship between big five personality traits, customer empowerment and customer satisfaction in the retail industry." Journal of Business and Retail Management Research (JBRMR) 11, no. 2 (2017).

26. PK Kanumarlapudi. “Improving Data Market Implementation Using Gray Relational Analysis in Decentralized Environments” Journal of Artificial intelligence and Machine Learning., 2024, vol. 2, no. 1, pp. 1–7. doi: https://dx.doi.org/10.55124/ jaim.v2i1.271
27. PK Kanumarlapudi. (2023) Strategic Assessment of Data Mesh Implementation in the Pharma Sector: An Edas-Based Decision-Making Approach. SOJ Mater Sci Eng 9(3): 1-9. DOI: 10.15226/2473-3032/9/3/00183
28. Raghavendra Sunku. (2025). Beyond digitalization strategic automation as a driver of policy administration performance using linear and random forest regression. International Journal of Computer Science and Data Engineering, 2(4), 260. https://doi.org/10.55124/csdb.v2i4.260