Strategic Customer Segmentation and Demand Prediction

Tushar Sharma[1], Kumar Saurav Jha[1], Jonathan Kartchman[1], Pradeep Somasundaram[1]
[1] Columbian College of Arts & Sciences, The George Washington University, Washington D.C. 20052, USA

Author Note

Correspondence concerning this article should be addressed to Tushar Sharma, Kumar Saurav Jha, Jonathan Kratchman, and Pradeep Somasundaram, Columbian College of Arts & Sciences, The George Washington University, Phillips Hall, 801 22nd St NW, Washington, DC 20052, United States.

1. Abstract

The strategic grouping of customers, known as customer segmentation, is a cornerstone of modern retail strategy, allowing companies to tailor their offerings and marketing efforts to different audience clusters. This study harnesses this concept, aiming to segment customers and predict demand by analyzing historical pricing data. We employ k-means and Hierarchical clustering, two clustering algorithms to categorize customers into distinct segments. This segmentation is based on purchasing behaviors and the revenue they help the company generate. Complementing this, we predict the daily and weekly demand which is in terms of the sales of different products using an XGBoost regression model. Our approach seeks to provide a nuanced understanding of customer dynamics and deliver accurate demand projections, which are critical for inventory management, pricing strategies, and personalized marketing. The anticipated outcome of our study is a practical framework that enables retailers to adapt to market changes proactively and to serve their customers more effectively.

Keywords: Customer Segmentation, Demand, E-commerce, Machine Learning, K-means, Hierarchical clustering, XGBoost

2. Introduction

In the dynamic landscape of contemporary e-commerce, consumer segmentation has emerged as a transformative paradigm, driven by sophisticated algorithms that harness the power of machine learning [1]. In the realm of e-commerce, customer segmentation is a pivotal strategy employed by businesses to better understand and cater to the diverse needs of their customer base. By categorizing customers into distinct segments based on their purchasing behaviors, preferences, and demographics, e-commerce platforms can tailor marketing strategies, product recommendations, and user experiences. For instance, one segment may consist of bargain hunters who respond well to discounts and promotions, while another may comprise high-value customers who prioritize premium products and personalized services. Customer segmentation in e-commerce enables targeted communication, personalized marketing campaigns, and the optimization of product offerings. Through the analysis of historical data, e-commerce businesses can identify patterns, enhance customer engagement, and ultimately improve overall customer satisfaction and

loyalty. These methods pave the way for personalized and responsive marketing strategies, wherein items are marketed at different prices to each segment.

Previous studies have explored unsupervised learning clustering techniques for customer segmentation [2, 3]. Some studies also compared deep learning and dimensionality reduction techniques for customer segmentation in the telecom industry [4]. This study delves into the core of customer segmentation within the e-commerce sector, focusing on the intricate application of machine learning techniques, including k-means and Hierarchical clustering. By exploring the potential of data mining techniques, the research seeks to unravel the complexities of customer segmentation, aiming to refine pricing decisions for a more targeted and nuanced approach. Additionally, the study investigates the role of predictive modeling, utilizing algorithms such as extreme gradient boosting (XGBoost), and emphasizes the importance of time series cross-validation in such types of data with temporal quality and significance.

The remainder of this paper is structured as follows: First, the data is described along with a quick analysis of data followed by the methodology for each technique used in this study. Then the results for customer segmentation and demand predictions are discussed. Finally, conclusions are drawn based on the results obtained.

3. Methodology

3.1 Exploratory Data Analysis

The data contains historical sales data from an anonymous retailer, including information on orders, customers, and products [5]. There were a total of 24 fields and 51291 records in the data to begin with. Most categorical variables such as Customer name, City, State, Country, Product ID, etc. have high cardinality (many unique records) and are therefore practically, not very helpful in model building and evaluations. Critical columns from the dataset retained for the analysis include Order Date, Ship Date, Ship Mode, Segment, Market, Region, Category, Sub-Category, Sales, Quantity, Discount, Profit, Shipping Cost, and Order Priority. This comprehensive data forms the backbone of the study, providing a rich source of information for mining and analysis.

The exploratory data analysis (EDA) phase, crucial for understanding the underlying structure of the dataset. The EDA includes numerical feature distribution analysis, which likely involves descriptive statistics and visualization techniques such as box-whisker plots to understand the distribution of key metrics like sales, profit, and shipping costs. Scatter plots help to identify relationships and patterns between different variables, such as sales and profit, or shipping cost and order quantity. The correlation matrix further quantifies these relationships, revealing, for example, a strong positive correlation between sales and shipping cost, and a moderate positive correlation between sales and profit. Such insights are valuable in understanding which features may play a significant role in customer segmentation and demand prediction. Fig. 1 shows the feature distribution using histograms and fig. 2 shows the correlation between the numerical features based on Pearson correlation method.

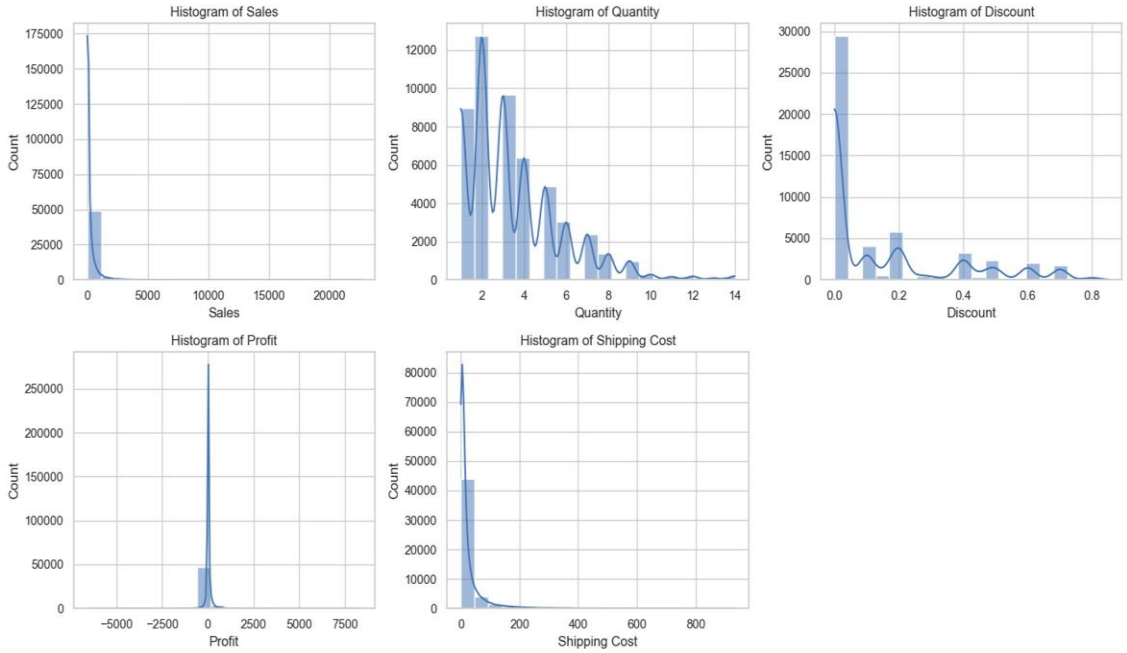

Fig. 1. Histograms showing the distribution of numeric features.

The histograms for feature distribution are highly skewed for most features. Particularly, the sales are mostly concentrated under the 1000 mark. Similarly, Profit seems to be concentrated around 0 which is a disadvantage of visualizing in this way because later in this study, certain customer segments show high profits while others do not. Shipping cost is also highly right skewed in its distribution. Histograms of this type of data may not be a suitable way of analyzing the data which is a collection of so many different categories of products, customers of different demographics.

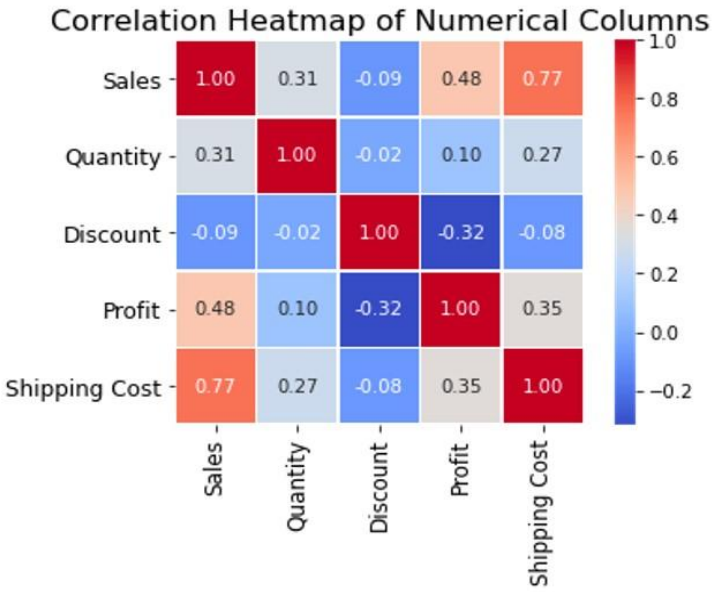

Fig. 2. Correlation matrix of numeric features based on Pearson correlation.

The correlation matrix shows that features such as Sales and Shipping Cost have a high positive correlation of 0.77. There is moderate positive correlation between Sales and Profit at 0.48. Discount and

Profit have a negative correlation of -0.32 which is kind of expected between such two quantities.

The evolution of the monthly (aggregated by different sub-categories) sales of different categories of products with time has been shown in fig. 3. The three major categories of products are Furniture, Office Supplies, and Technology. It can be observed that the monthly sales increase with time for all the three categories of products with periodic peaks and troughs in the trend.

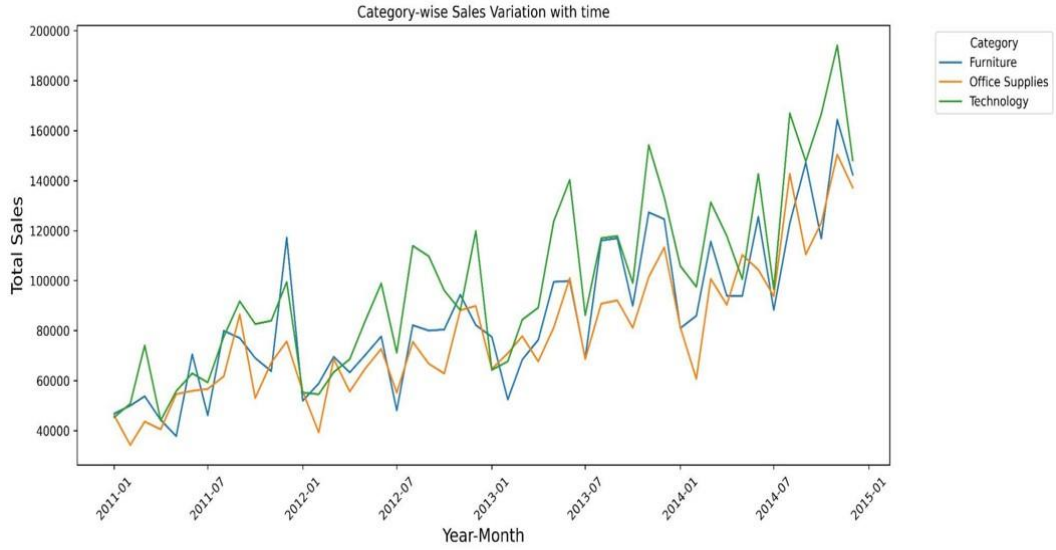

Fig. 3. Monthly sales of three categories of products over time.

3.2 Customer Segmentation

3.2.1 k-means Clustering

K-Means clustering is employed in this project to categorize customers into distinct segments based on their purchasing behavior and their impact on the company's growth. The fundamental idea behind k-means clustering is to group data points into 'k' clusters, where each cluster represents a set of similar observations [6]. The algorithm operates through an iterative process that begins by randomly selecting k initial cluster centroids, where k represents the desired number of clusters. It then assigns each data point to the cluster whose centroid is closest, typically measured using Euclidean distance. Subsequently, the centroids are recalculated as the mean of the points within each cluster. This assignment and update process is repeated iteratively until convergence, where the centroids stabilize, and data points no longer switch clusters. The algorithm aims to minimize the variance within clusters and maximize the separation between them. The optimal ‘k’ is determined by the elbow-method which is based on the Within-Cluster-Sum-Square (WCSS) metric given as:

$$\mathrm{WCSS} = \sum_{i \in n} (x_i - y_i)^2 \tag{1}$$

where, ‘y_i’ is centroid for observation ‘x_i’ and ’n’ is the total number of observations.

By applying this technique to customer data, the project aims to uncover inherent patterns and preferences within the customer base, facilitating a more tailored approach to demand predictions.

3.2.2 Hierarchical Clustering

Hierarchical clustering is a method of grouping data points into hierarchical structures, known as dendrograms, based on their similarity [7]. The algorithm begins by treating each data point as an individual cluster and then iteratively merging the closest clusters until only one cluster remains. The closeness or dissimilarity between clusters is measured using various metrics, such as Euclidean distance or correlation coefficients. Hierarchical clustering can be agglomerative, starting with individual data points and progressively merging them, or divisive, beginning with a single cluster and recursively splitting it. The result is a tree-like structure that visually represents the relationships and hierarchical organization of the data. One advantage of hierarchical clustering is its ability to provide a detailed representation of the data's structure at different levels of granularity. However, its computational complexity can be higher compared to other clustering methods.

While k-means provides a straightforward, fixed-number partitioning, hierarchical clustering offers flexibility, revealing the data's natural hierarchy without requiring a predetermined cluster count. The choice between these methods often hinges on the desired level of granularity, interpretability, and the inherent structure of the dataset.

3.3 Demand Prediction

3.3.1 XGBoost Regression

XGBoost, short for Extreme Gradient Boosting, is an advanced implementation of the gradient boosting algorithm that has gained widespread popularity in both machine learning competitions and practical applications [8]. As a decision-tree-based ensemble learning method, XGBoost excels in regression, classification, ranking, and custom prediction tasks. Its strength lies in constructing a series of decision trees sequentially, with each new tree correcting errors from previous ones. The ‘boosting’ aspect involves iterative improvement, enhancing overall performance, while the ‘gradient’ component uses the gradient descent algorithm to minimize loss when adding new models. The efficiency and effectiveness of XGBoost make it a preferred choice for diverse predictive modeling challenges.

In the context of the project, XGBoost is used for demand prediction based on historical data. The algorithm can handle various types of predictive modeling tasks required in the project, such as regression tasks for predicting continuous variables like sales and classification tasks for determining customer segments. Due to its ability to manage large datasets and work with sparse data, XGBoost is well-suited for time series data, which can be vast and have many types of features, including numeric, categorical, and time-engineered.

Metrics like Mean Absolute Percentage Error (MAPE) and Root Mean Square Percentage Error (RMSPE) are used in this study to check how accurate the prediction models are. The MAPE calculates the average percentage difference between predicted and actual values, while RMSPE gives a normalized measure of the difference between what we predicted and what actually happened but it emphasizes the squared differences. These metrics are crucial for knowing how precise our daily demand predictions are. These metrics are given as:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|a_i - f_i|}{f_i} \times 100 \tag{2}$$

$$\text{RMSPE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{a_i - f_i}{f_i}\right)^2} \times 100 \tag{3}$$

where, n is the total number of records, a_i is the actual value of the record, and f_i is the predicted/forecasted value of that record.

For best model performance, hyperparameter tuning has been performed in each customer segment using the GridsearchCV algorithm from Sci-kit Learn. The hyperparameter search space for the XGBoost model is given in table 1.

Number of trees	Max tree depth	Learning rate
100	1	0.001
200	3	0.01
500	5	0.05
1000	10	0.1

Table 1. Hyperparameter search space for the XGBoost model.

3.3.2. Time Series Cross Validation

Time Series Cross-Validation is a technique used to evaluate the predictive performance of a time series model. Unlike standard cross-validation methods, which assume that the data points are independent and identically distributed, time series data are inherently ordered by their timestamp, and their values are often serially correlated. This correlation means that random splitting, a common approach in cross-validation for non-temporal data, is inappropriate for time series because it could lead to significant leakage of information from the future into the past, thereby producing overly optimistic and unreliable performance estimates. In time series cross-validation, the data are split into a series of training and testing sets over time. Each split involves a training set that includes all data up to a certain point in time, and a testing set that includes data following that point. As the validation process iterates, the training set increases in size, incorporating more data points, and the testing set shifts forward in time. This approach respects the temporal order of observations, ensuring that the model is always validated on data that occur after the training data.

In this study, for demand predictions, 5-fold time series cross validation technique has been employed wherein the data of each identified customer segment is five training-testing splits (folds). The training set has a size that is given by the following expression:

$$size_{training} = \frac{i \times n_{samples}}{n_{splits} + 1} + n_{samples} \bmod\left(n_{splits} + 1\right) \tag{4}$$

where, i is the current split, *n_samples* is the total data size, and *n_splits* is the total number of cross validation splits.

The time series cross validation folds for a segment have been shown in fig. 4.

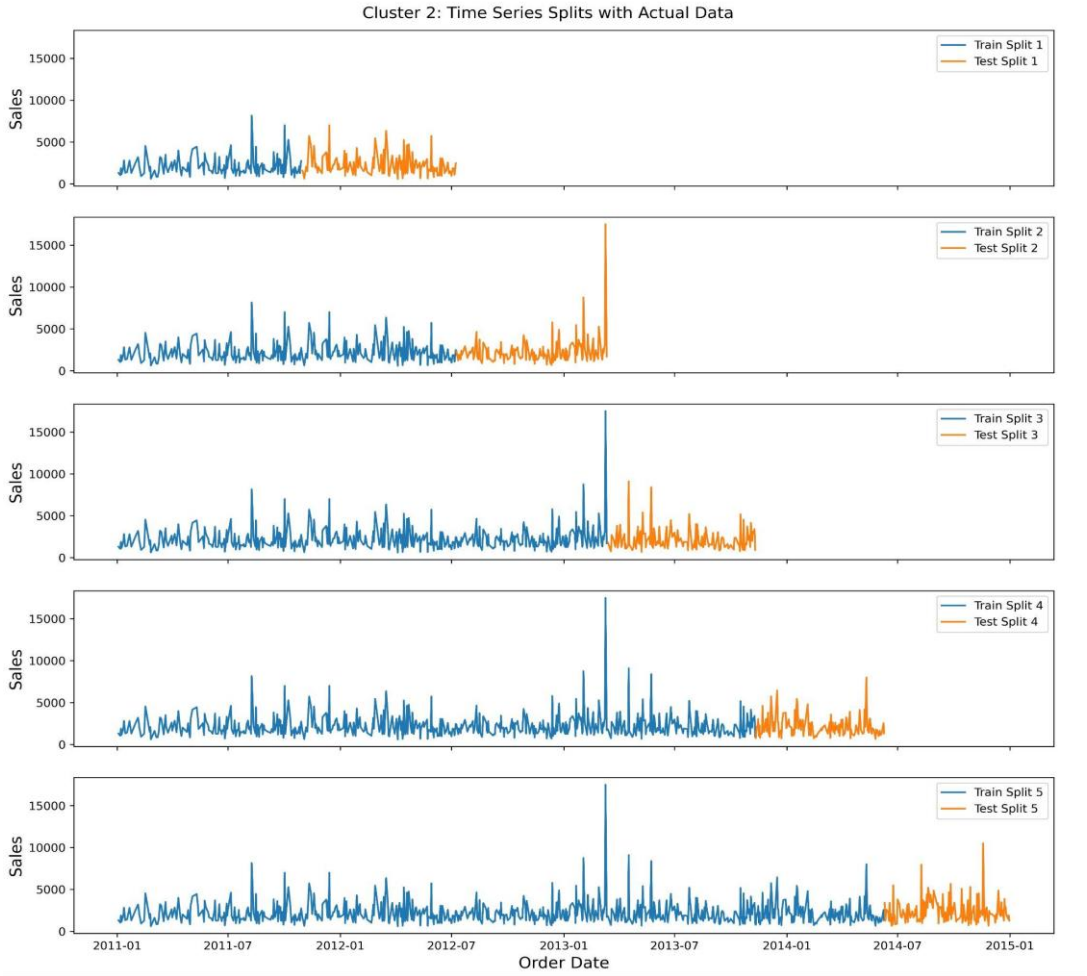

Fig. 4. Time series cross validation folds for segment 2. The training data iteratively grows in size as the testing data from the previous fold is added to training data in the subsequent fold.

4. Results

4.1 Customer Segmentation (k-means vs Hierarchical clustering)

Based on the WCSS method, there were four optimal number of segments (k = 4) retained from k-means clustering as shown in fig. 5. The corresponding customer segments (clusters) obtained from k-means clustering are also shown in fig. 5. The four identified segments based on profit and sales were labeled as 'Stable Performers' (segment 1), 'High Achievers' (segment 2), 'Challenged Margins' (segment 3), and 'Balanced Growth' (segment 4). The results for the insights from the final customer segments between k-means clustering have been summarized in table 2. For Hierarchical clustering, the number of clusters were experimented taking three and four clusters. For the sample data experiment, the Silhouette score for Hierarchical clustering came out to be 0.737 and that for k-means clustering came out to be 0.686. A comparison of results from k-means clustering and Hierarchical clustering for the sample data has been summarized in table 2.

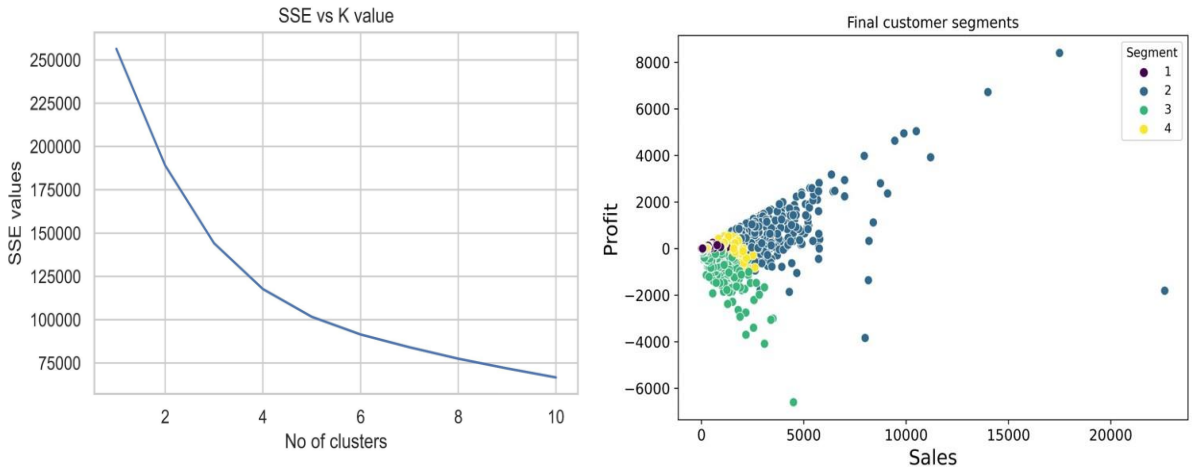

Fig. 5. WCSS method for optimal number of clusters (left) and final customer segments (right) from k-means clustering.

Parameter	Hierarchical	k-means
Optimal no. of clusters	3, 4	4
Silhouette Score	0.737	0.686
Cluster 1 size	2201	2041
Cluster 2 size	43	106
Cluster 3 size	304	391
Cluster 4 size	16	26

Table 2. Comparison of results of k-means and Hierarchical clustering techniques for customer segmentation.

Though the Silhouette score was higher for Hierarchical clustering, the distribution of data was more or less comparable to that of k-means clustering. Furthermore, there was an insignificant difference between the Silhouette score for Hierarchical clustering with three and four clusters. These observations coupled with the fact that Hierarchical clustering is difficult to implement on the entire data with 51000 records is the reason why the final customer segments retained were taken from k-means clustering. From the final customer segment insights in table 3., segment 2 shows the highest average sales and profit. The average sales in this segment is much higher than the rest of the segments. On the other hand, segment 2 does not generate any profitable margin for the company as indicated by the negative value of profit (-75.73). This suggests that segment 2 should be targeted strategically to increase profit whereas, more concern and care should be taken while marketing in segment 3. Segment 4 shows moderate sales and profit and segment 1 shows very less sales and profit.

Segment	Sales	Quantity (avg/head)	Discount	Profit	Shipping cost
1	128.93	2.28	0.04	26.15	13.31
2	2075.61	5.90	0.05	466.12	247.74
3	131.31	3.05	0.52	-75.73	13.81
4	352.76	6.30	0.06	58.43	35.29

Table 3. Insights in terms of average parameters of the four customer segments identified from k-means clustering.

4.2 Demand Prediction

4.2.1 Daily demand

In predicting daily demand, the XGBoost regression model was implemented, utilizing a set of retained features including numeric variables, select categorical variables such as ship mode and order priority, and engineered time features like year, month, and day. The modeling approach, as discussed before, employed time series cross-validation with 5 folds, segmented within each of the four customer segments. While generating the results, the 5 folds were combined on a single time series for better visualization and the metrics were also averaged across the folds. The results for daily demand predictions suggest that segment 3 shows the highest average MAPE and RMSPE of 43.34 and 170.08, respectively which is rather high. On the other hand, segment 2 had the most accurate predictions with an average MAPE and RMSPE of 22.04 and 28.66, respectively. Overall, daily demand is fairly accurately predicted by the XGBoost model advocating its usefulness in time series predictions. Fig. 6. shows the results for the daily demand across the four customer segments.

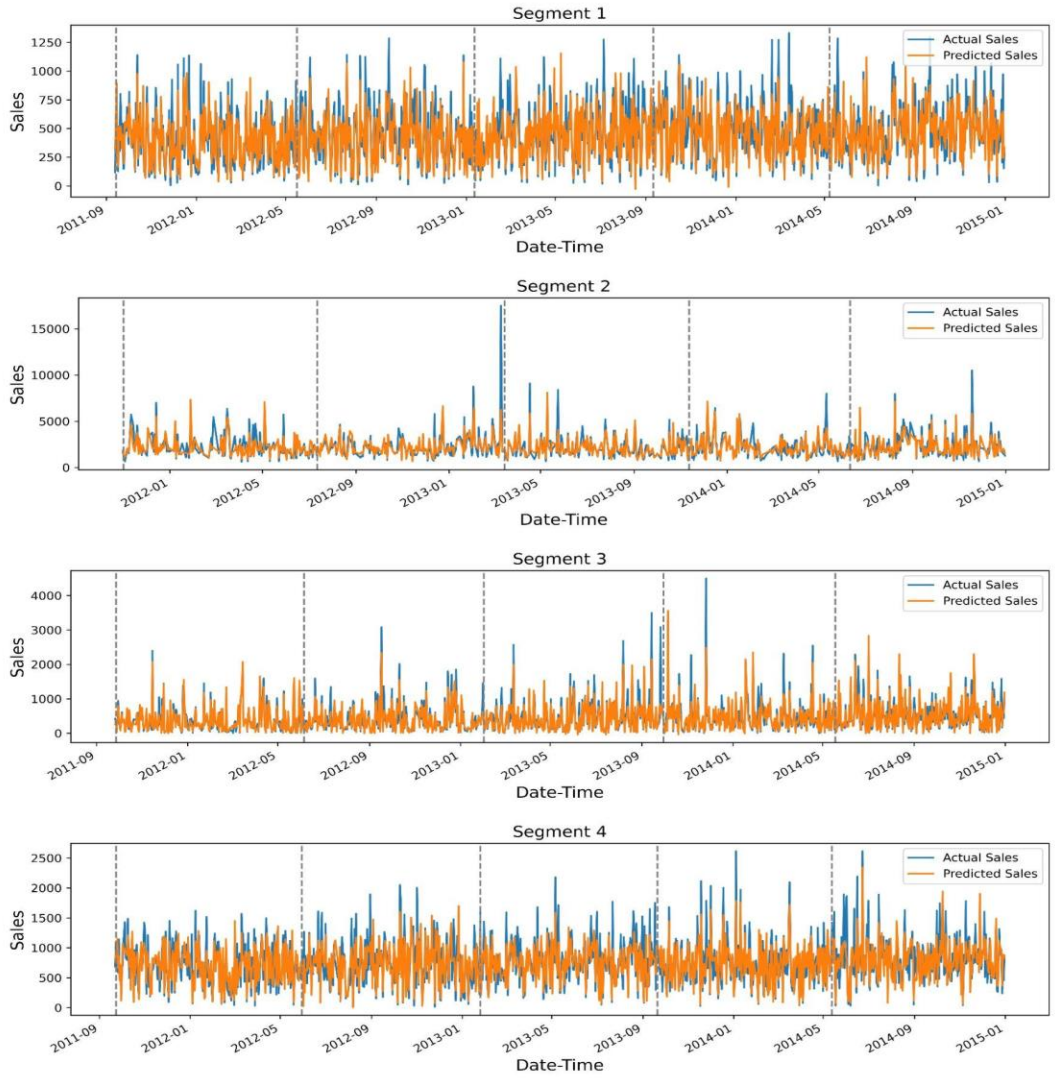

Fig. 6. Actual vs predicted daily demand (sales) for the four customer segments by XGBoost model. The dashed lines divide the 5 cross validation folds.

The feature importance was calculated for all features retained and the results for each segment are shown in fig. 7. It can be observed that features like 'Shipping cost' and 'Order priority' are the most important across all segments for predicting daily demand. These are followed by 'Profit' and 'Ship mode' and 'Discount' and 'Quantity' have moderate importance. It is important to note that time-engineered features (day, day of week, month, year) turned out to be not very important in predicting daily demand. This could perhaps be because of the irregular purchasing pattern of the products of different categories with respect to time.

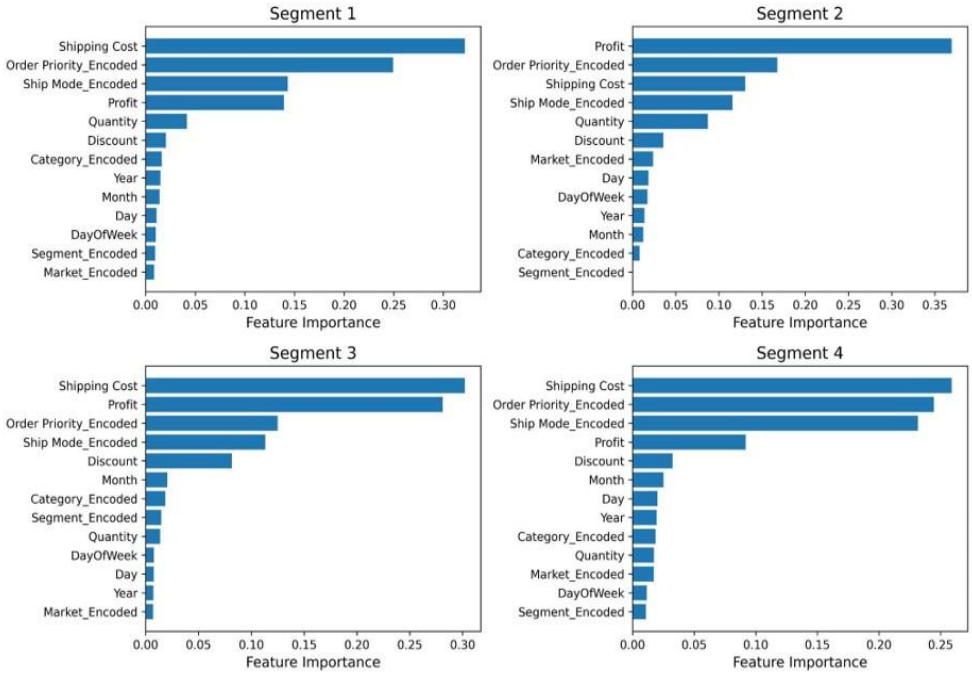

Fig. 7. Feature importance barplots for daily demand predictions in all customer segments.

4.2.2 Weekly demand

For the prediction of weekly demand, the sales were aggregated over the week and specific features such as quantity, profit, shipping cost, and discount were retained. The aggregation process involved summing up weekly sales, profit, and quantity, whereas, taking the mean for discount and shipping cost as summing these quantities would not be technically meaningful. This time, the time engineered features are dropped because they were found ineffective in predictions. Also, no categorical variables were retained for the purpose of predicting weekly demand as their aggregation is not possible the same way as for the numerical features. This methodology aimed to capture and leverage key metrics at a broader weekly level, providing a comprehensive perspective on demand trends over time. The results for weekly demand are shown in fig. 8. The results indicate that segment 1 had the best weekly demand predictions with an average MAPE and RMSPE of 8.50 and 10.57, respectively. As opposed to daily demand where segment 3 had the weakest accuracy in terms of predictions, for weekly demand, segment 2 has the highest prediction errors. This is in direct contrast because segment 2 had the best predictions for daily demand. A plausible explanation for this could be that customers in segment 2 purchase more frequently aiding in more data records each day. However, aggregating the data to a weekly timestamp would cause a reduction in the data effectively used for model training and testing. Overall, the weekly demand predictions are more accurate

compared to daily demand suggesting it may be suitable to perform aggregated predictions and forecasting in such problems.

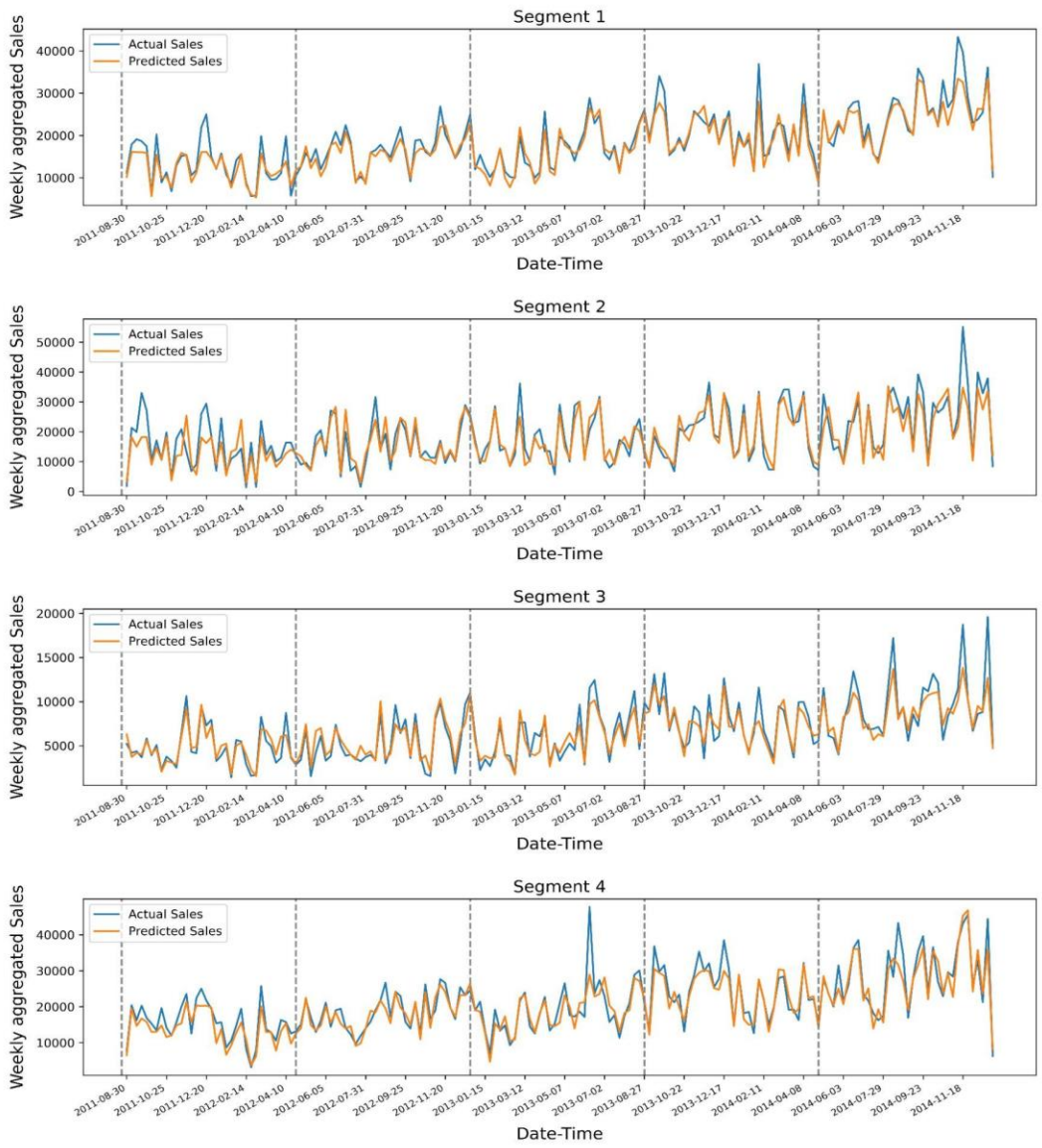

Fig. 8. Actual vs predicted weekly demand (sales) for the four customer segments by XGBoost model. The dashed lines divide the 5 cross validation folds.

The feature importance for weekly demand as shown in fig. 9. suggests that ‘Quantity’ of the products stands out an important predictor across all segments. Profit is also an important predictor. ‘Shipping cost’ and ‘Discount’ do not carry much importance for weekly demand predictions. Thus, only two features are effectively able to predict the weekly demand across all segments. However, addition of features could probably improve the model predictions given the number of features currently is so less.

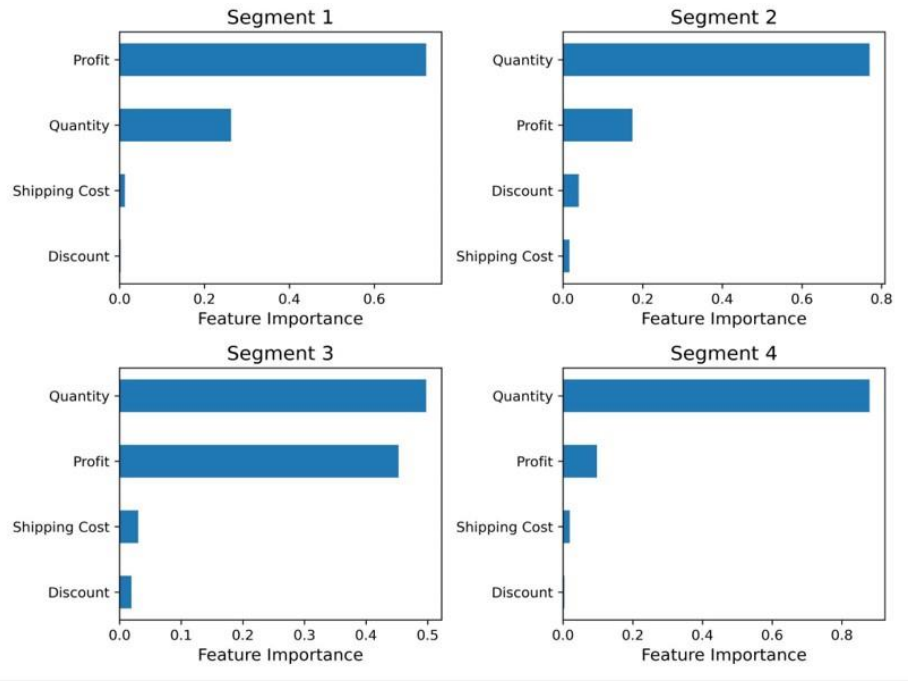

Fig. 9. Feature importance barplots for weekly demand predictions in all customer segments.

5. Conclusions

This project set out to explore the dynamic and complex nature of customer behavior within the retail domain, aiming to provide a granular understanding through customer segmentation and to enhance business operations via accurate demand predictions in each of the identified segments. Practical considerations led to the selection of k-means clustering for its ease of implementation on large datasets. The identified segments—'Stable Performers,' 'High Achievers,' 'Challenged Margins,' and 'Balanced Growth'—provide nuanced perspectives on customer behaviors and revenue contributions.

For demand predictions, the XGBoost model demonstrated efficacy in predicting both daily and weekly sales across the identified customer segments, although given the distribution of data, the weekly demand was predicted with greater accuracy as shown by the model metrics. Notably, the model revealed varying prediction accuracy among segments, emphasizing the importance of tailoring forecasting approaches to distinct customer behaviors.

Overall, the study's findings present a practical framework for retailers, aiding in proactive adaptation to market changes, effective inventory management, and personalized marketing strategies. The demonstrated success of the XGBoost model in predicting demand underscores its utility in retail analytics, providing a valuable tool for optimizing pricing strategies and enhancing customer-centric decision-making.

References

1. Policarpo, L. M., da Silveira, D. E., da Rosa Righi, R., Stoffel, R. A., da Costa, C. A., Barbosa, J. L. V., ... & Arcot, T. (2021). Machine learning through the lens of e-commerce initiatives: An up-to-date systematic literature review. *Computer Science Review*, *41*, 100414.

2. Kansal, T., Bahuguna, S., Singh, V., & Choudhury, T. (2018, December). Customer segmentation using K-means clustering. In *2018 international conference on computational techniques, electronics and mechanical systems (CTEMS)* (pp. 135-139). IEEE.

3. Smeureanu, I., Ruxanda, G., & Badea, L. M. (2013). Customer segmentation in private banking sector using machine learning techniques. *Journal of Business Economics and Management*, *14*(5), 923-939.

4. Alkhayrat, M., Aljnidi, M., & Aljoumaa, K. (2020). A comparative dimensionality reduction study in telecom customer segmentation using deep learning and PCA. *Journal of Big Data*, *7*, 1-23.

5. https://www.kaggle.com/datasets/apoorvaappz/global-super-store-dataset

6. MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, No. 14, pp. 281-297).

7. Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, *32*(3), 241-254.

8. Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).