# Strategic Customer Segmentation and Demand Prediction

Jonathan Kratchman

Kumar Saurav Jha

Tushar Sharma

Pradeep Somasundaram

Introduction to Data Mining, The George Washington University

# Author Note

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

# 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. Our project harnesses this concept, aiming to segment customers and forecast demand by analyzing historical pricing data. We employ K-Mean Clustering, an algorithm adept at discovering patterns within data, to categorize customers of an undisclosed retail store into distinct segments. This segmentation is based on purchasing behaviors, providing a granular view of the customer base. Complementing this, we implement XGBoost regression, a machine learning technique renowned for its predictive strength. This model anticipates future demand by analyzing historical data trends while adjusting for economic variables such as inflation. Our approach seeks to provide a nuanced understanding of customer dynamics and to 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 Forecasting, Machine Learning ,Predictive Modeling*

# Introduction

Consumer segmentation represents a paradigm shift in the contemporary e-commerce landscape, a sophisticated algorithm-driven strategy that divides customers into groups and leads to marketing items at different prices to each segment. This method promises enhanced profitability for businesses while purporting to offer value for consumers, creating a more responsive and personalized shopping experience.

This study dives into the heart of consumer segmentation, exploring the intricacies of machine learning applications that refine pricing decisions within the e-commerce sector. Firstly, it investigates the potential of data mining techniques to segment customers into distinct categories, thereby enabling more nuanced and targeted pricing strategies. It interrogates the role of predictive modeling and

ization in establishing pricing structures that not only elevate revenue but also resonate with diverse customer segments.

Ultimately, the intention is to furnish stakeholders with actionable insights that leverage the full spectrum of data-driven pricing strategies, delivering a robust model that augments business acumen with predictive precision.

## Objectives

In a digital marketplace where consumer choices are abundant, the strategic application of machine learning to pricing decisions represents a significant competitive advantage. This study explores the potent combination of data mining and predictive modeling to refine e-commerce pricing strategies. We delve into customer segmentation using data mining techniques, aiming to discern distinct customer groups characterized by unique purchasing behaviors, demographics, and preferences. This categorization is fundamental for businesses to tailor their marketing efforts and pricing schemes.

# Methodology

The methodology of customer segmentation through data mining and the predictive modeling for optimal pricing includes two specific clustering techniques, K-means and hierarchical clustering, are compared for their effectiveness in segmenting customers. This comparison is crucial as different clustering methods can yield varied results in terms of the granularity and accuracy of the segments created. K-means clustering is an unsupervised learning approach utilized here to segment customers based on their impact on the company's profitability and sales. The method involves partitioning the data into k distinct clusters by minimizing within-cluster variances and is often visualized using the within-cluster sum of squares (WCSS) to determine the optimal number of clusters through the elbow method.

Hierarchical clustering, on the other hand, provides higher interpretability through dendrograms that illustrate the hierarchy of cluster formation and is considered a more naturalistic and flexible approach, albeit less efficient with large datasets. A silhouette score, which measures how similar an object is to its own cluster compared to other clusters, is used to validate the number of clusters chosen.

"Stable Performers" with moderate sales and quantity, low discounts, positive profits, and moderate shipping costs. "High Achievers" characterized by high sales and quantities, moderate discounts, high profits, and high shipping costs. "Challenged Margins" exhibiting low sales and moderate quantities, high discounts, negative profits, and moderate shipping costs. "Balanced Growth" with moderate to high sales and quantities, low to moderate discounts, positive profits, and moderate shipping costs.

These segments reveal the diversity within the customer base, reflecting varying levels of profitability and sales performance. The identification of these segments is essential for developing tailored marketing strategies and personalized pricing models that can cater to the specific needs and behaviors of different customer groups.

The predictive modeling for demand prediction, where factors such as shipment costs, order priority, and profit have emerged as significant predictors for daily demand, while order quantity and profit are crucial for weekly demand predictions. These insights are pivotal for businesses to forecast demand accurately, manage inventory efficiently, and set prices that can adapt to market changes and consumer behavior patterns. We use time as our primary feature to forecast future demand more effectively. This could provide a substantial benefit to the company by optimizing pricing strategies and improving decision-making processes based on predictive analytics.

### Exploratory Data Analysis

The datasets contain historical sales data from an anonymous retailer, including information on orders, customers, and products. Critical columns from the dataset retained for analysis include Order ID, Order Date, Ship Date, Ship Mode, Customer ID, Customer Name, Segment, City, State, Country, Postal Code, Market, Region, Product ID, Category, Sub-Category, Product Name, 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.

The culmination of the EDA: The identification of customer segments through k-means clustering and the evaluation of cluster validity through silhouette scores. The silhouette score results, which measure the cohesion within clusters and the separation between them, ensuring that the chosen clustering is meaningful and distinct. The number of clusters and their respective sizes would also be discussed, providing insight into the customer base's segmentation.

The researchers are laying the groundwork for a comprehensive analysis. They're establishing the data's credibility, understanding its characteristics, and beginning to unlock the actionable insights hidden within through statistical and machine learning techniques. The resulting customer segments form the basis for developing tailored pricing strategies and demand forecasts that are sensitive to customer preferences and behavior, which is the next step in the research process.

### K-Means Clustering

This approach involves retaining numeric and selected categorical features such as ship mode and order priority, along with engineered time features like year, month, and day. This implies that the model considers both the type of shipping and the urgency of the order, as well as temporal trends in its predictions. The methodology indicates the use of time series cross-validation with multiple folds within each customer segment, which ensures the robustness and generalizability of the predictive model. Furthermore, the utilization of an XGBoost model for this purpose, with a focus on 5-fold cross-validation for training and testing, as well as hyperparameter tuning to optimize the model's performance.

The metrics such as the Mean Absolute Percentage Error (MAPE) and Root Mean Square Percentage Error (RMSPE), which are standard for evaluating the accuracy of forecasting models. MAPE measures the size of the error in percentage terms, and RMSPE provides a normalized measure of the deviation between the predicted and actual values. These metrics are essential for understanding the precision of the demand forecasts daily.

The weekly demand prediction. In this context, features such as quantity, profit, shipping cost, and discount are aggregated on a weekly basis. The process of summing up weekly sales, profit, and quantity, while calculating the weekly mean of discount and shipping costs. This aggregation likely serves as the basis for a different predictive model tailored to capture the more extended trends in the data.

Like the daily demand prediction, the weekly demand forecast model would also employ MAPE and RMSPE to evaluate its accuracy. These metrics are particularly important in the context of inventory management and pricing strategies, as they indicate the reliability of the model on a longer time scale, which is crucial for strategic planning and resource allocation in e-commerce.

In summarizing, it's evident that the research has transitioned from identifying customer segments to leveraging those segments for refined demand forecasting. The application of XGBoost, a powerful gradient boosting algorithm, for both daily and weekly demand predictions, demonstrates a sophisticated approach to understanding and anticipating market needs. The rigorous validation methods and the attention to error metrics underscore the scientific rigor of the study, aiming to provide actionable insights to predict demand.

### Hierarchical Clustering

The predictive analytics aspect of the research, specifically regarding the importance of various features in predicting daily and weekly demand. The results of the feature importance analysis, which is a critical step in understanding which factors have the most significant impact on demand patterns.

For daily demand prediction, factors such as shipment cost, order priority, and profit might have been identified as the most influential predictors. This implies that the urgency of an order and the cost associated with shipping it significantly affects the number of daily orders. The importance of profit as a predictor suggests that more profitable items could potentially see higher daily demand, likely due to better customer satisfaction or higher value perception.

For weekly demand, the order quantity and profit are the key predictors. The inclusion of order quantity highlights the potential for bulk orders or larger transactions to influence the overall weekly demand, while profit remains a consistent predictor, reinforcing the importance of profitability in demand forecasting.

Feature importance analysis is crucial as it helps in refining the model by focusing on the most impactful variables, potentially reducing complexity and improving prediction accuracy. It is also an insightful process for business strategy, as it pinpoints where the company should concentrate its efforts to boost demand. For instance, if shipment cost is a significant predictor, optimizing shipping logistics might be a key area for business development.

The potential improvements for the predictive models. Advanced modeling techniques could include ensemble methods or deep learning algorithms that might capture complex nonlinear relationships in the data more effectively. The inclusion of lag features refers to using data from previous time periods as predictors for current demand, recognizing patterns over time, which could improve the model's ability to forecast future demand.

In summarizing, it becomes evident that the research has reached a sophisticated level of analysis, where predictive models are scrutinized for the effectiveness of their features in forecasting demand. The insights gathered from this stage are instrumental for the retailer to understand the levers of customer demand and adjust their inventory and pricing strategies accordingly.

### XGBoost Regression

XGBoost, which stands for Extreme Gradient Boosting, is an advanced and efficient implementation of the gradient boosting algorithm. It has gained immense popularity in machine learning competitions and practical applications due to its performance and speed. XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In predictive modeling challenges, it can be utilized for regression, classification, ranking, and user-defined prediction problems.

At its core, XGBoost is an ensemble learning method that constructs a series of decision trees in a sequential manner. Each new tree corrects the errors made by the previous ones. The "boosting" aspect of the method refers to this iterative improvement, as each new model incrementally improves upon its predecessor, hence "boosting" the overall performance. The "gradient" part pertains to the use of gradient descent algorithm to minimize the loss when adding new models.

One of the reasons for XGBoost's effectiveness is that it employs a regularized model formalization to control over-fitting, which provides it with better performance. This regularization feature is a significant distinction from traditional gradient boosting. Moreover, XGBoost offers several advanced features for model tuning, including handling missing data, tree pruning, and built-in cross-validation. The algorithm can automatically learn the best imputation value for missing data while training, which is a substantial advantage over other algorithms that require pre-processing to handle such scenarios.

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 e-commerce data, which can be vast and have many categorical features.

Another significant advantage of XGBoost is its scalability and efficiency. It is designed to be computationally efficient by utilizing both hardware and software optimizations, such as cache awareness and block structure for parallel computation. It can run on a single machine as well as a distributed framework, making it versatile for different scales of data.

XGBoost also includes a feature importance metric, which is critical for understanding the factors that drive the prediction. In the project, knowing which variables most significantly affect customer demand can inform business decisions regarding pricing strategies, inventory management, and marketing efforts.

In conclusion, XGBoost's robustness, versatility, and performance make it a strong candidate for the project's predictive modeling needs. It provides a way to capture complex non-linear relationships within the data, which are essential for accurate demand forecasting. Furthermore, its ability to handle various types of data, its efficiency in computation, and its model tuning capabilities align well with the project's goals of segmenting customers and predicting demand within an e-commerce setting.

### 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, also known as time-based 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.

One common approach to time series cross-validation is the "rolling forecast" method, where the test set consists of a single future time point and the model is retrained before each prediction is made. Another variant is the "expanding window" approach, where the size of the test set remains fixed, but the training set grows with each fold, including all available data up until the beginning of the test set.

Time series cross-validation is particularly important for projects like the one described, where demand forecasting is key. In such a scenario, models need to be tested on unseen data to ensure that they can accurately predict future demand. This method allows researchers to estimate the performance of predictive models in a way that is more aligned with how the models will be used in practice.

The careful splitting of data in time series cross-validation also allows for the detection of temporal patterns, such as seasonality and trends, and how these patterns affect the model's predictions. For example, if demand for a product increase during the holiday season, a model trained on data that includes several holiday seasons and validated through time series cross-validation is more likely to capture and predict this seasonal increase accurately.

Time series cross-validation is a robust method for evaluating the performance of time series models, providing a realistic assessment of how well a model will perform in making future predictions. This technique is crucial for any project where the temporal dimension cannot be ignored and where the integrity of the time series data must be maintained throughout the model evaluation process.

# Outcomes

To initiate with a deeper discussion into the demand prediction models, delineating both the daily and weekly forecasting techniques. For daily demand forecasting, they might include details on the selected machine learning algorithms, their performance metrics, and the significance of various features in predicting demand. One would expect to see a comparison of different models, their Mean Absolute Percentage Error (MAPE), and Root Mean Squared Percentage Error (RMSPE), providing insights into the accuracy and reliability of the models.

For weekly demand forecasting, the outline the process of aggregating data on a weekly basis, discussing how features such as total sales, profit, shipping cost, and discount rates contribute to demand patterns over longer periods. The methodology for aggregating data, the nuances of weekly versus daily forecasting, and the implications of these different time frames on inventory management and pricing strategies would be crucial elements of this discussion.

In terms of results, the outcomes of the demand prediction model, highlighting which features are most influential in driving demand. This could include visual representations like feature importance charts, which help to convey the relative importance of variables such as shipment cost, order priority, profit, and order quantity in the models. Such insights are vital for understanding how different aspects of the business impact customer purchasing behavior and for informing strategic decisions.

Additionally, the model might address model improvements and future work. For instance, the inclusion of more sophisticated modeling techniques, like deep learning, could be suggested for capturing complex patterns in the data. The use of lag features—historical data points used to predict future events—might also be proposed to enhance the model's predictive power.

The conclusion would synthesize the findings, reflecting on the implications for the retail store's strategic planning. Key takeaways would include the identification of customer segments with the highest and lowest margins, how this segmentation can inform targeted marketing strategies, and the potential of future dynamic pricing models to optimize revenue.

Overall, these would likely encapsulate the core achievements of the project, the analytical rigor of the methodologies employed, and the actionable insights derived from the research. The conclusion would emphasize the practical applications of the study, suggesting ways in which the retailer can apply these insights to improve business outcomes.

# Conclusion

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 forecasting. Through the meticulous application of K-Mean Clustering and XGBoost regression models, we have successfully categorized customers into distinct groups and predicted future demand by accounting for various influential factors, including the overarching impact of inflation.

The segmentation of customers based on purchasing behavior, demographics, and preferences has yielded significant insights, highlighting the diversity within the customer base and underscoring the potential for targeted marketing strategies. The identification of segments such as "Stable Performers," "High Achievers," and "Balanced Growth" customers allows for personalized engagement and strategic resource allocation, ensuring that the retailer can meet the nuanced demands of each group effectively.

The use of XGBoost regression models has been particularly enlightening, showcasing the algorithm's robustness in handling large datasets and its efficiency in generating reliable demand forecasts. This project underscores the value of adopting advanced machine learning techniques in the retail sector, offering a scalable solution that can adapt to changing market conditions.

As we reflect on the achievements of this research, we are reminded of the importance of continuous improvement. While our models have performed admirably, there remains room for further refinement. Future work could explore the integration of additional predictive features, the implementation of more sophisticated machine learning algorithms, and the extension of our models to encompass real-time data feeds for even more agile demand forecasting.

In conclusion, the findings from this project provide a compelling case for the use of data analytics and machine learning in e-commerce. By leveraging these tools, retailers can gain a deeper understanding of their customers, anticipate market trends, and make informed decisions that drive profitability and growth. This project not only contributes to academic knowledge in the field but also offers practical, data-driven strategies that can be directly applied within the industry.

# References

R. Gupta, A. K. Yadav, S. Jha and P. K. Pathak, "Time Series Forecasting of Solar Power Generation Using Facebook Prophet and XG Boost," 2022 IEEE Delhi Section Conference (DELCON), New Delhi, India, 2022, pp. 1-5

[doi: 10.1109/DELCON54057.2022.9752916](https://ieeexplore.ieee.org/document/9752916)