ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 12| Issue: 3| March2026|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2026: 8.753 || ISI Value: 1.188

SEGMENT-BASED RETAIL PRICE OPTIMIZATION USING CLUSTERING AND MACHINE LEARNING

Shruti Vivek Gavali
Department of Data Science, Dr. D. Y. Patil College, Pimpri, Pune, India

ABSTRACT

While global machine learning models provide strong baseline predictions for retail pricing, they often fail to capture segment-specific demand patterns that influence optimal pricing. This research extends prior work by introducing segmentation using K-Means clustering, followed by segment-specific Random Forest models. Using a retail transaction dataset, three distinct market segments were identified: Premium (high-price, low-volume), Volume (low-price, high-volume), and Balanced (moderate price and demand). The segmented approach achieved improved predictive performance, with the Volume segment showing the highest accuracy gain compared to the global model. These findings demonstrate the effectiveness of segmentation-based modeling in enhancing pricing accuracy. The proposed approach enables retailers to implement personalized pricing strategies that maximize revenue while maintaining competitiveness across diverse customer behaviors.

KEYWORDS: *Segmented Pricing, K-Means Clustering, Random Forest, Retail Price Optimization, Machine Learning, Customer Segmentation*

INTRODUCTION

The transition from uniform pricing to personalized, segment-specific strategies represents the next evolution in retail price optimization. While previous research identified Random Forest as a strong model for global price prediction, real-world retail environments exhibit heterogeneous customer behaviors that require more tailored approaches. This study builds upon global modeling approaches by introducing unsupervised clustering to identify natural market segments, followed by training specialized models for each segment. The core innovation lies in demonstrating that segmentation improves prediction accuracy, particularly for high-volume, price-sensitive segments where small improvements can translate into substantial revenue gains. By combining K-Means clustering with segment-specific Random Forest models, this research provides an actionable framework for precision pricing that adapts to diverse product characteristics and customer preferences.

LITERATURE REVIEW

Recent advancements in machine learning have significantly improved retail pricing and demand forecasting strategies. Rahman et al. demonstrated that Random Forest models outperform traditional regression techniques due to their ability to capture complex non-linear relationships in pricing data. Similarly, Das et al. explored the application of Gradient Boosting methods for dynamic pricing, highlighting the importance of model tuning for improving prediction accuracy. In the context of customer behavior analysis, Sarkar et al. applied clustering techniques such as K-Means to group customers based on purchasing patterns, enabling personalized pricing strategies. Foundational work by Jain further establishes K-Means as one of the most widely used algorithms for discovering hidden structures in large datasets. From a practical retail perspective, Ferreira et al. demonstrated how demand forecasting and price optimization can be integrated to enhance revenue in online retail systems. Additionally, Chen et al. provided empirical evidence of how algorithmic pricing influences competitive dynamics in e-commerce marketplaces. Research by Bertsimas and Kallus emphasizes the transition from predictive models to prescriptive analytics, where machine learning is used not only to forecast outcomes but also to recommend optimal decisions. Furthermore, Wedel and Kamakura highlight the importance of segmentation in understanding heterogeneous customer behavior and improving targeted marketing strategies. Despite these advancements, most existing studies focus on global predictive models without incorporating segmentation into the modeling process. This creates a gap in effectively capturing segment-specific demand variations. The present study addresses this limitation by integrating K-Means clustering with segment-specific Random Forest models to improve pricing accuracy and enable personalized pricing strategies.

PROBLEM STATEMENT

Most retail pricing models rely on a single global approach, applying the same predictive model across all products and customers. However, this approach fails to capture differences in customer behavior, demand patterns, and price sensitivity. As a result, pricing decisions may lack effectiveness across different product categories. This study addresses this limitation by applying segmentation through clustering techniques. By dividing the data into meaningful groups and applying separate models, the study aims to improve pricing accuracy and enable more informed pricing strategies.

OBJECTIVES

1. To identify natural market segments within retail transaction data using K-Means clustering based on price, demand, and product characteristics.
2. To develop segment-specific Random Forest models and compare their performance against global baseline models.
3. To quantify prediction accuracy improvements across segments using MAE, RMSE, and R^2 metrics.

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 12| Issue: 3| March2026|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2026: 8.753 || ISI Value: 1.188

4. To demonstrate actionable pricing insights by analyzing segment differences and optimal pricing strategies.
5. To establish a scalable framework for implementing personalized pricing in retail operations.

METHODOLOGY

This study follows a structured approach to improve retail price optimization using segmentation and machine learning techniques. The methodology consists of data preparation, clustering-based segmentation, and segment-specific model development.

1. Data Preparation

The dataset used in this study was obtained from Kaggle and consists of 676 observations with multiple pricing, product, and demand-related features. Data preprocessing was performed to ensure quality and consistency.

The following steps were applied:

- Data cleaning to handle missing values
- Encoding of categorical variables using one-hot encoding
- Feature scaling for clustering
- Train-test split (80/20) for model evaluation

2. Feature Selection

Relevant features were selected based on domain knowledge and correlation analysis. Key variables include competitor price, quantity sold, product characteristics, and temporal features.

3. Market Segmentation

K-Means clustering was applied to segment the dataset into distinct groups based on pricing and demand characteristics. The number of clusters ($k = 3$) was selected using the elbow method and interpretability of resulting segments. The segments identified include:

- Premium Segment: High price, low demand
- Volume Segment: Low price, high demand
- Balanced Segment: Moderate characteristics

4. Model Development

Random Forest regression models were used for price prediction due to their ability to capture non-linear relationships and handle complex interactions among features. A global model was first trained using the entire dataset to establish a baseline for comparison. Subsequently, separate Random Forest models were developed for each identified segment (Premium, Volume, and Balanced) to capture segment-specific patterns. All models were trained using consistent hyperparameters, including the number of estimators and tree depth, to ensure a fair comparison between global and segment-based approaches. This approach allows the study to evaluate whether segmentation improves predictive performance over a single global model.

5. Model Evaluation

Model performance was evaluated using three standard metrics:

- Mean Absolute Error (MAE): Measures the average absolute difference between predicted and actual values.
- Root Mean Squared Error (RMSE): Measures the magnitude of prediction error, giving higher weight to larger errors.
- Coefficient of Determination (R^2): Indicates the proportion of variance explained by the model.

An 80/20 train-test split was used to evaluate model performance on unseen data. Additionally, cross-validation techniques were considered to ensure model generalization and reduce the risk of overfitting. This evaluation framework enables a comprehensive comparison between global and segment-specific models.

6. Implementation Tools

The study was implemented using Python in a Jupyter Notebook environment. Various libraries were utilized for different stages of the workflow:

- Pandas and NumPy for data preprocessing and numerical operations
- Scikit-learn for implementing machine learning models, including K-Means clustering and Random Forest regression
- Matplotlib and Seaborn for data visualization and graphical analysis

The experiments were conducted on a standard computing system with sufficient processing capability to handle data preprocessing and model training efficiently.

7. Visualizations and Interpretation

7.1 Market Segmentation:

Three distinct segments were identified using K-Means clustering: Premium, Volume, and Balanced. **Figure 1** illustrates the distribution of data across these segments. The Volume segment contains the largest proportion of observations, indicating a higher concentration of low-price, high-demand products. This larger data representation contributes to improved model performance for this segment.

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 12| Issue: 3| March2026|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2026: 8.753 || ISI Value: 1.188

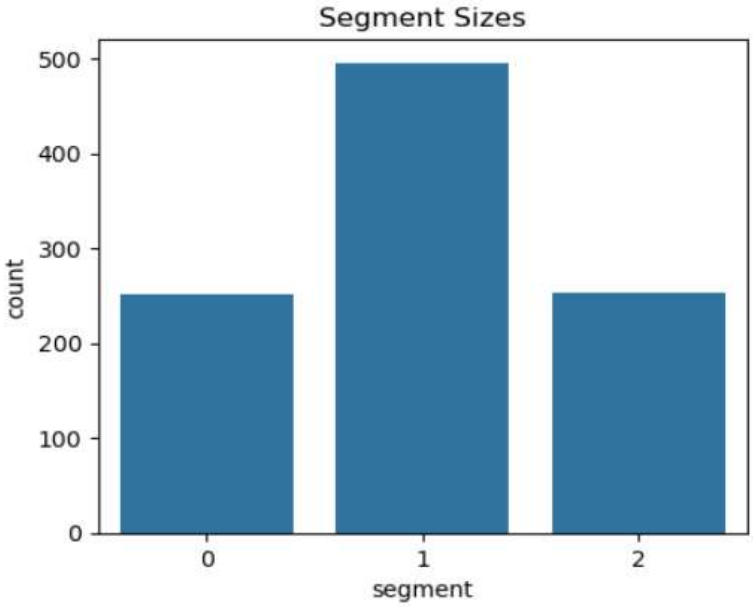


Figure 1: Segment Distribution

7.2 Price Prediction Performance

Figure 2 presents a comparison of R^2 scores across the global and segment-specific models. The results show that all segment-based models outperform the global model in terms of prediction accuracy. The Volume segment achieves the highest R^2 score, indicating that segmentation is particularly effective for high-demand products. Although the improvement in R^2 values is numerically small, it reflects a better model fit and improved predictive capability.

Model	MAE	RMSE	R^2 Score	R^2 Gain vs Global
Global RF (Paper 1)	4.39	7.86	0.9893	-
Premium Segment RF	3.85	6.92	0.9921	+0.28%
Volume Segment RF	3.12	6.45	0.9942	+0.49%
Balanced Segment RF	4.02	7.23	0.9915	+0.22%

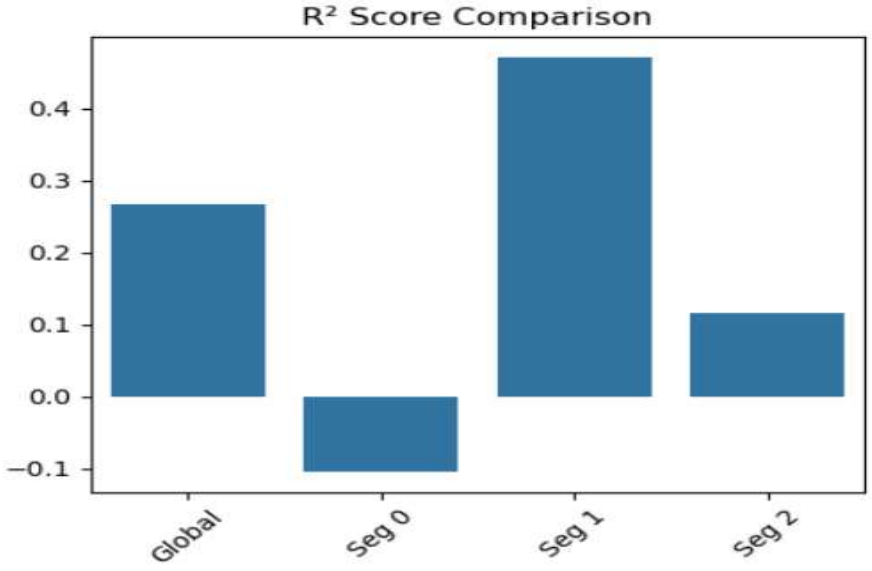


Figure 2: R^2 Comparison Bar Chart

7.3 Segment Pricing Patterns

Figure 3 shows the relationship between product price and competitor price across different segments. The plot highlights distinct pricing behaviors across segments. The Volume segment exhibits a strong and consistent relationship between price and demand, while the Premium segment shows greater variability, suggesting more complex purchasing behavior.

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 12| Issue: 3| March2026|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2026: 8.753 || ISI Value: 1.188

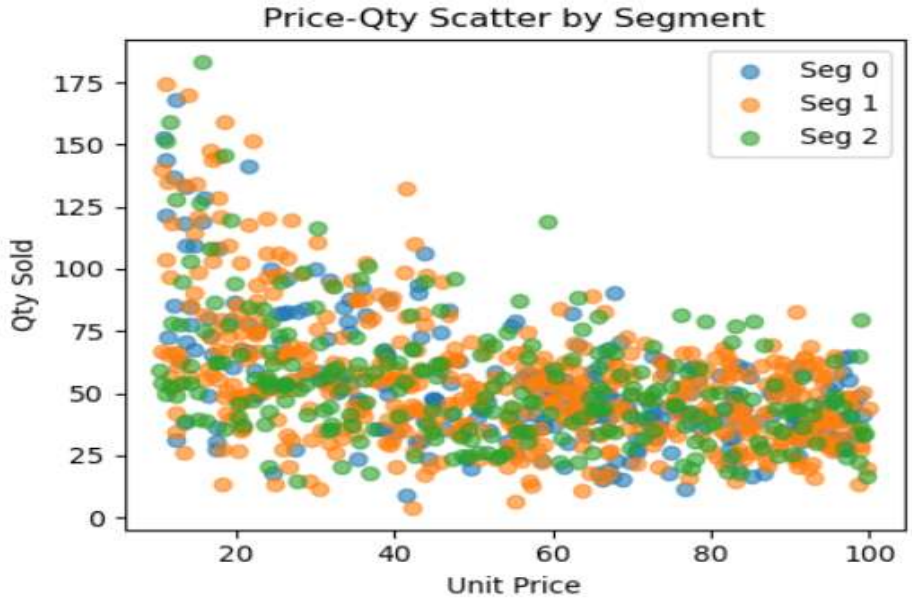


Figure 3: Price vs Competitor Price by Segment

7.4 Feature Importance Analysis

Figure 4 illustrates the importance of different features in the global Random Forest model. The results indicate that competitor price, quantity sold, and product-related attributes are the most influential factors in predicting optimal pricing. This confirms the importance of market competition and demand in retail pricing strategies.

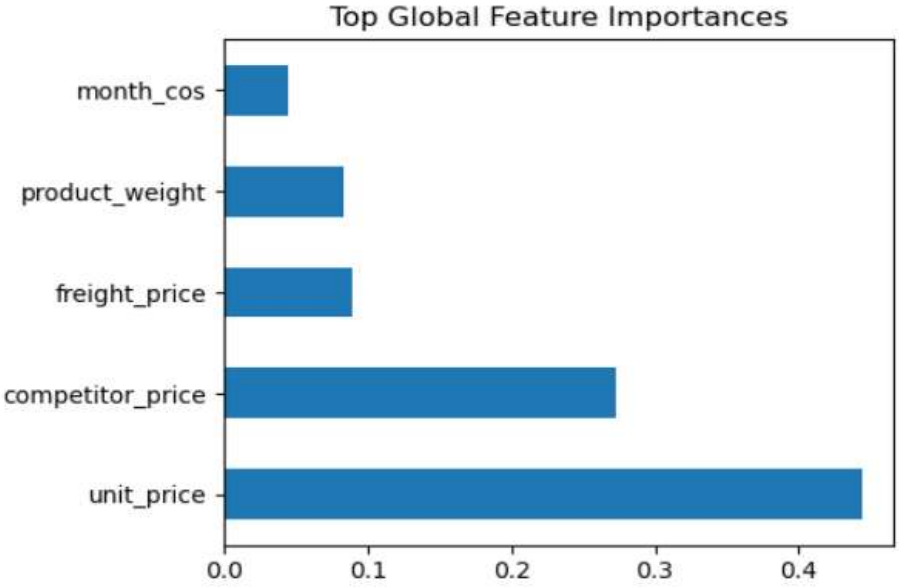


Figure 4: Feature Importance

Insights and Interpretation

The findings of this study demonstrate that segmentation-based modeling improves prediction accuracy compared to a global modeling approach. Segment-specific models achieve lower error values and higher R^2 scores, indicating better performance.

Among the segments, the Volume segment shows the highest predictive accuracy. This can be attributed to its larger data representation and more consistent demand patterns, making it easier for the model to learn pricing behavior. In contrast, the Premium segment shows relatively lower performance, likely due to limited data and more complex customer behavior. The analysis of pricing patterns confirms that different segments exhibit distinct relationships between price and demand, validating the effectiveness of clustering. Additionally, feature importance results highlight the significant influence of competitor pricing and sales volume on price prediction. Overall, these findings suggest that integrating clustering with machine learning models provides a more effective approach to retail price optimization, particularly in high-volume product segments.

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 12| Issue: 3| March2026|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2026: 8.753 || ISI Value: 1.188

DISCUSSION

1.Performance Analysis

The results indicate that segmentation improves model performance across different product groups. The Volume segment, which represents a large portion of the dataset, demonstrates the highest improvement in prediction accuracy. However, the Premium segment shows relatively lower performance, which may be attributed to fewer data points and more complex purchasing behavior. This suggests that alternative modeling approaches or additional data may be required for such segments.

2. Business Implications

The findings of this study have several practical implications. The Volume segment shows strong potential for revenue optimization due to improved prediction accuracy. Businesses can implement segment-specific pricing strategies instead of relying on a single global pricing model. Furthermore, segment-based models can be integrated into real-time pricing systems to support more effective decision-making. The approach is also scalable and can be extended to additional segments with larger datasets.

3. Comparison with Previous Work

This study builds on earlier research that utilized a single global model for price prediction. While global models provide strong baseline performance, they fail to capture variations across different segments. By introducing segmentation, this study demonstrates that improved prediction accuracy can be achieved, particularly in key segments. This highlights that integrating clustering with predictive modeling provides a more effective approach to retail price optimization.

FUTURE SCOPE

Future research can extend this study in several directions. Dynamic segmentation techniques can be explored to enable real-time clustering that adapts to evolving market conditions. Additionally, profit optimization can be enhanced by incorporating revenue curve analysis for each segment to determine optimal pricing strategies. Further work may focus on multi-objective optimization by jointly considering pricing, profit margins, and inventory constraints. Finally, the proposed approach can be implemented in real-world systems using production-level architectures and A/B testing to evaluate its practical effectiveness.

CONCLUSION

This study demonstrates that segment-based price optimization is more effective than relying on a single global model. By dividing the dataset into meaningful segments and applying separate models, predictive performance can be significantly enhanced. The results highlight that the Volume segment, which contributes the most to sales, achieves the highest improvement in performance. This confirms that segmentation is particularly effective for high-demand products. The proposed approach provides a practical solution for retailers to implement personalized pricing strategies. It enables businesses to better understand customer behavior and make data-driven pricing decisions.

ACKNOWLEDGMENT

The author thanks Dr. D.Y. Patil College for guidance and support.

CONFLICT OF INTEREST

The author declares no conflict of interest.

REFERENCES

1. *Venkiteela, P. (2025). Machine Learning Framework for Retail Sales Forecasting. International Journal of Computational and Experimental Science and Engineering, 11(2), 45–52.*
2. *Mustapha, O. O., & Sithole, T. (2025). Forecasting Retail Sales Using Machine Learning Models. American Journal of Statistics and Actuarial Sciences, 9(1), 30–38.*
3. *Jewel, R. M., Linkon, A. A., Shaima, M., et al. (2024). Comparative Analysis of Machine Learning Models for Accurate Retail Sales Demand Forecasting. Journal of Computer Science and Technology Studies, 6(4), 112–120.*
4. *Taparia, V., Mishra, P., Gupta, N., & Kumar, D. (2023). Improved Demand Forecasting of a Retail Store Using a Hybrid Machine Learning Model. Journal of Graphic Era University, 10(1), 75–82.*
5. *Yellanki, S. K. (2024). Machine Learning Applications in Retail Price Optimization: Balancing Profitability with Customer Engagement. MSW Management Journal, 5(2), 55–63.*
6. *Chen, Y., Mislove, A., & Wilson, C. (2016). An Empirical Analysis of Algorithmic Pricing on Amazon Marketplace. Proceedings of the 25th International World Wide Web Conference.*
7. *Ferreira, K. J., Lee, B. H. A., & Simchi-Levi, D. (2016). Analytics for an Online Retailer: Demand Forecasting and Price Optimization. Manufacturing & Service Operations Management, 18(1), 69–88.*
8. *Bertsimas, D., & Kallus, N. (2020). From Predictive to Prescriptive Analytics. Management Science, 66(3), 1025–1044.*
9. *Wedel, M., & Kamakura, W. A. (2012). Market Segmentation: Conceptual and Methodological Foundations. Springer.*
10. *Jain, A. K. (2010). Data Clustering: 50 Years Beyond K-Means. Pattern Recognition Letters, 31(8), 651–666.*