

Customer Segmentation

Yuxuan Zhang

Westcliff University

AIT 506

Instructor: Professor Noha Bhairy

Date: Nov 22, 2025

Customer Segmentation

For the jupyter notebook, see <https://github.com/yuki172/ait506-class-assignments/tree/main/4> .

GAP does not allow me to upload a jupyter notebook.

The goal of this project is to perform customer segmentation, interpret the results and use the results to recommend marketing strategies for the relevant organization to help it improve sales and increase revenue. The dataset used is the Customer Shopping Trends Dataset on kaggle.

For preprocessing, I use one hot encoding for all categorical variables and used Z-score normalization to standardize the variables (de Amorim, 2023). For clustering, I used k-means clustering. More specifically, I used the silhouette score to determine the optimal number of clusters and performed k-means clustering (Rousseeuw, 1987).

For interpretation, I used the optimal number of clusters as indicated by the silhouette scores to compute k means. I then converted the cluster centroids back to the original scale for interpretation. The clusters show clear characteristics that can be used to generate business decisions to improve sales and revenue. The first cluster can be characterized by a high purchase amount and high previous purchase amount. Moreover, most of the products bought by this cluster are boots, which indicates strong product loyalty. For these customers, we can make marketing plans centered around boots and footwear (Oliver, 1999). For example, we can open early exclusive access and discounts. We can also recommend items that go with boots such as belts and jackets. The marketing effort spent on these customers should be strongest since they are more likely to invest in the products that are recommended.

The second cluster consists of customers who are average shoppers and have broad interests in different products. Due to the lack of a clear favorite product, it might be worthwhile

to incorporate recommendation tools to personalize the shopping experience. This way, if the recommendations are right, these customers can be more satisfied with the shopping experience which will lead to more sales. The third cluster is characterized by the fact that most of the purchase is for coats and that the purchase is sporadic. It is hard to make a marketing plan for these customers because they already know what they want and when they want them. Therefore, improving the coats themselves might be a more worthwhile effort.

References

- de Amorim, L. B. V., Cavalcanti, G. D. C., & Cruz, R. M. O. (2023). The choice of scaling technique matters for classification performance. *Applied Soft Computing*, 133, 109924. <https://doi.org/10.1016/j.asoc.2022.109924>
- Oliver, R. L. (1999). *Whence consumer loyalty?* *Journal of Marketing*, 63(Special Issue), 33–44. <https://doi.org/10.2307/1252099>
- Rousseeuw, P. J. (1987). *Silhouettes: A graphical aid to the interpretation and validation of cluster analysis*. *Journal of Computational and Applied Mathematics*, 20, 53–65. [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7)