Customer Segmentation Clustering based on Demographics and Behaviors

Spencer Hirsch

April 25, 2024

1 Abstract

Identifying similaries between customers can be a useful strategy for companies when creating marketing campaigns. By analyzing the demographics and the purchasing behaviors of customers, a company can gain better insight into their customer base. With valuable insight there is potential to release marketing campaigns targeted towards these groups to increase sales. By employing unsupervised machine learning we can better identify the characteristics of similar groups to aid in such campaigns. This study aims to utilize customer demographic and purchasing behaviors to identify trends and group consumers together based on these characteristics. Through the use of three different unsupervised machine learning clustering methods we can identify groups with shared characteristics and attempt to optimize these groups with a variety of preprocessing and feature reduction methods to increase the effectiveness of these models in an attempt to find the most effective model for this study.

2 Introduction

3 Methods

In order to prepare the data for this exploration, it was necessary to clean up the data to adhere to the task at hand. Due to the nature of the project it was fitting that marketing campaign was removed from the dataset, the initial dataset contained 5 marketing campaigns as well as response information from the customers. For the prupose of clustering based on demographic and behaviors this information was not necessarily important.

In addition to this feature removal, much of the existing data in the dataset was categorical, rather than leaving objects in the dataframe, the data was converted to contain only numeric values, this process included 1-hot encoding all categorical data and splitting the datetime object into three new features, month, day, and year. Following the manual deletion and modification of the original dataset, the new dataset consisted of 33 features.

Feature Name	Feature Description
Id	Unique customer identifier.
Year_Birth	Birth year of customer.
Education	Highest level of education obtained by customer.
Marital_Status	Marital status of customer.
Income	Annual income of customer.
Kidhome	Number of young children in the home.
Teenhome	Number of teenagers in the home.
Dt_Customer	Date when customer first enrolled.
Recency	Last visit of customer.
MntWines	Amount spent on wines.
MntFruits	Amount spent on fruits.
MntMeatProducts	Amount spent on meats.
MntFishProducts	Amount spent on fish.
MntSweetProducts	Amount spent on sweets.
MntGoldProds	Amount spent on gold products.
NumDealsPurchases	Number of purchases made as part of a discount promotion.
NumWebPurchases	Number of purchases made through website.
NumCatalogPurchases	Number of purchases made through catalog
NumStorePurchases	Number of purchases made in store.
NumWebVisitsMonth	Number of times customer has visited the website.

Table 1: Original columns in dataset after removing marketing campaign information.

3.1 Preprocessing

After manual feature reduction and cleaning was performed, additional preprocessing was done on the data. This included utilizing a variety of methods: Standard Scaler, Robust Scaler, Quantile Transform, and Log Transform. All of these methods were used in an attempt to increase the effectiveness of the clusters. We can see that some of the preprocessing methods formed stonger clusters than others. Robust Scaler, seen in figure two, was the worst preprocessing used when it came to clustering. The method that had the strongest clustering was the preprocessing performed with Quantile Transformation, seen in Figure 3.

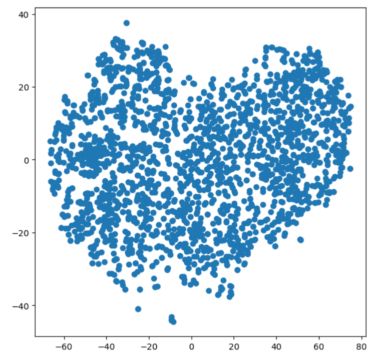

Figure 2: Visual representation of data preprocessed using Robust Scaler.

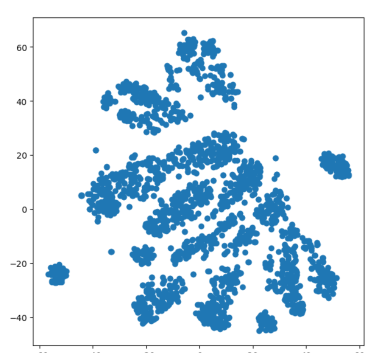

Figure 1: Visual representation of data preprocessed using Standard Scaler.

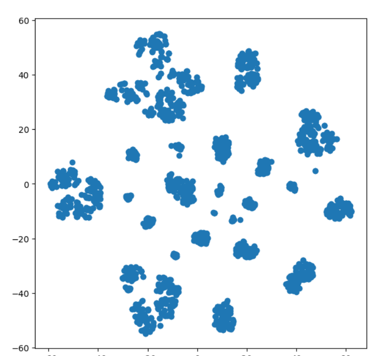

Figure 3: Visual representation of data preprocessed using Quantile Transform.

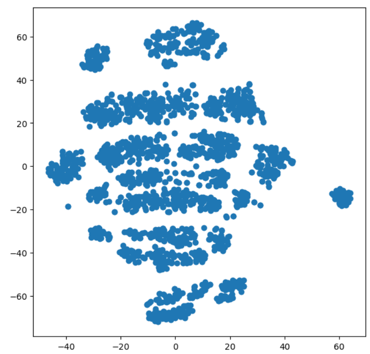

Figure 4: Visual representation of data preprocessed using Log Transformed.

3.2 Feature Reduction

After the dataset had gone through manual reduction

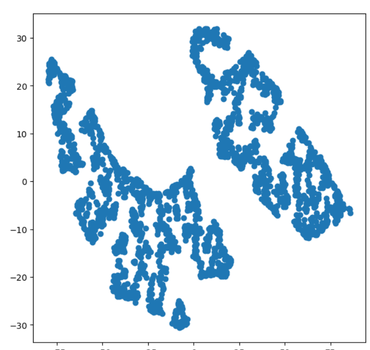

Figure 5: Visual representation of data with t-SNE feature reduction with 2 n_clusters.

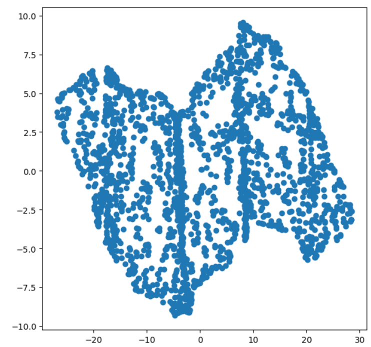

Figure 6: Visual representation of data with t-SNE feature reduction with 3 n_clusters.

3.3 Clustering

3.3.1 K-Means

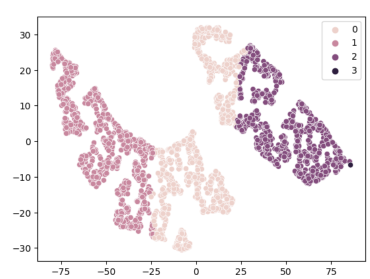

Figure 7: Visual representation of dataset clustered with K-Means method with t-SNE feature reduction applied.

3.3.2 Hierarchical

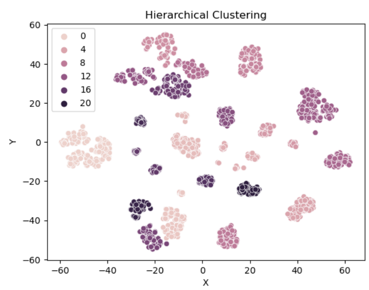

Figure 8: Visual representation of dataset clustered with Hierarchical method with t-SNE feature reduction and Quantile Transformation applied.

3.3.3 DBSCAN

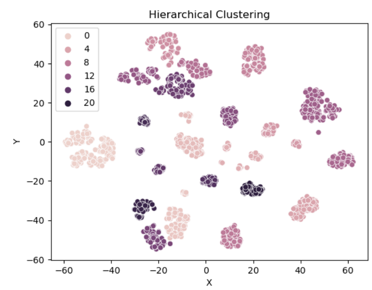

Figure 9: Visual representation of dataset clustered with DBSCAN method with t-SNE feature reduction and Quantile Transformation applied.

4 Results

5 Conclusion

References

[1] Patel, Vishakh. *Customer Segmentation/Clustering*. Kaggle. Accessed on April 22, 2024. Available online: `https://www.kaggle.com/datasets/vishakhdapat/customer-segmentation-clustering/data`

[2] Zhao, J. (2024). Customer segmentation application based on K-Means. *Applied and Computational Engineering, 47*, 242-247. `https://doi.org/10.54254/2755-2721/47/20241400`

[3] Hua, N., Leu, B. A., & Kumar, R. (2022). Machine learning-based customer segmentation. International Journal of Information Security, 9.

[4] John, Jeen & Shobayo, Olamilekan & Ogunleye, Bayode. (2023). An Exploration of Clustering Algorithms for Customer Segmentation in the UK Retail Market. Analytics. 2. 809-823. 10.3390/analytics2040042.

[5] Alie, Juhaini & Gustriansyah, Rendra. (2024). Customer Segmentation for Digital Marketing Based on Shopping Patterns. Jurnal Aplikasi Bisnis dan Manajemen. 10. 209-216. 10.17358/jabm.10.1.209.

[6] Li, Xiaotong & Lee, Young. (2024). Customer Segmentation Marketing Strategy Based on Big Data Analysis and Clustering Algorithm. Journal of Cases on Information Technology. 26. 1-16. 10.4018/JCIT.336916.