Customer Segmentation Clustering based on Demographics and Behaviors

Spencer Hirsch

April 25, 2024

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.

Keywords – Machine learning; customer segmentation; clustering algorithm

1 Introduction

Customer segmentation can be a powerful technique used by retailers in order to gain a better understanding of their target markets. For a business to be sucessful they must be able to cater towards their customer base, and address their specific needs. In the modern world, machine learning techniques can be utilized to make it easier for companies to better understand their customers. Customer segmentation can be used to break a larger group of people into smaller groups, separatiing them based on their shared characteristics, such as, income or how much an individual spents on wine. This information can be especially useful for marketing strategies as companies would be able to share special benefits with these groups of people to draw them in.

The data used for this exploration can be seen in Table 1, all data refers strictly to their demographic information or their shopping behaviors.[1] This dataset was not explicitly stated to be from any sort of market or company, however, it is fair to assume that the data was collected from some sort of grocery store. This seems to be the case for all features, the only feature that seems to be odd for a grocery store is the "Amount spent on gold products". Nonetheless, the other features appear to refer to a grocery store, therefore for the purpose of this study that is what we will assume.

1.1 Related Work

Customer segmentation can serve as a useful tool in many disciplines, Zhao dicusses a handful of potential uses when it comes to the application of customer segmentation. Whether it be for marketing purposes as previously mentioned but also for exploration purposes. Zhao brings up the possibility of utilizing this tool to determine new markets to enter.[2] This could be an incredibly useful tool for a company looking to expand the scope of their business. If a company idenities trends in their data indication that the market is starting to lean a specific direction, they would be able to tailor their products or services to meet this demand. Zhao's discussion of customer segmentation doesn't necessarily focus on strictly unsupervised machine learning methods, the article discusses two methods, recency, frequency, and monetary (RFM) and K-Mean clustering algorithms.[2] Rather than performing an exploration into the methods the article provides a general overview of the field at the time of publishing.

Rather that exploring potential methods of customer segmentation, Holy and Sokol bring customer segmentation into the real-world. Their study deals with the study and development of a clustering algorithm to be utilized by a major Czech drugstore chain.[3] Holy and Sokol utilize a variety of clustering methods paried with K-Means in order to better understand the patterns of customers, RFM, as discussed in Zhao,[2] purchased products structure (PPS), and store mission (SM).

Similarly to the study that I have conducted here, John, Shobayo, and Ogunleye aim to study a variety of unsupervised machine learning methods, along with Principle Component Analysis and RFM.[4] The goal of the study is similar to mine, as they are aiming to find the most effective method for customer segmentation. The authors of this paper explore, K-Means clustering, Gaussian mixture model (GMM), density-based spatial clistering of applications with noise (DBSCAN), agglomerative clustering, and balanced iterative reduction and clustering using hierarchies (BIRCH).[4] The study proved to perform well with the authors being able to acheieve a Silhoutette Score, a measure of consistency in the cluster, of 0.8, through the use of PCA and GMM.[4]

Just as the previous study, Hicham and Karim utilize a number of unsupervised machine learning models for the purposes of customer segmentation. Their paper proposes a clustering ensemble method of, DBSCAN, K-Means, Mini Batch K-Means, and Mean-Shift. Through the use of their method, they were able to acheive a Silhouette Score of 0.73.[5] This method proved to be useful, DBSCAN itself was only able to acheive 0.72 itself, so there was marginal gain with their technique.[5]

Turkmen also performs a study using a variety of methods in hopes of acheving greater results, in their paper they weight K-means clustering, hierarchical clustering, DBSCAN, and RFM.[6] The results of this study were not as great as the previous study using an ensemble method, Turkmen was able to acheive a Silhouette Score of 0.6 using the K-Means clustering algorithm.[6]

We can see that there have been numerous efforts and resources dedicated to studying the use of a variety of clustering methods for the purpose of customer segmentation. In the following study, we will see the reoccurence of clustering methods, such as, K-Means clustering, Hierarchical clustering, and DBSCAN clustering as well as the introduction of new preprocessing methods that were not used in the above studies.

2 Methods

Three methods of unsupervised machine learning clusering methods were used, K-Means Clustering, Hierarchical Clustering, and DBSCAN. In addition to these three methods a variety of preprocessing methods and feature reduction methods were used to reduce the dimensionality of the data. The goal of course to find the best combination of methods that provide the most efficient model. Just as John, Shobayo, and Ogunleye[4], Hicham and Karim[5], and Turkmen[6], the primary metric used in determing the sucess of the models was the Silhouette Score. This value helps us determine how well the clustering methods is constructing the clusters. A Silhouette Score closer to 1 signifies that the clusters are more closely associated, whereas a score close to -1 means that they are not.

Once accessing the dataset, we start by manually preprocessing the data, this includes removing some of the features in the dataset as well modifying categorical features of the dataset. Once this was complete, preprocessing using a variety of algorithms was employed, as well as feature reduction algorithms.

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 removal of marketing campaign information.

Once this was complete, the three clustering models where employed and underwent hyper-parameter tuning to find the optimal number of clusters for the highest Silhouette Score.

2.1 Preprocessing

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 five 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.

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 Transformation, and Log Transformation. 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 method used when it came to clustering. The method that had the strongest clustering was the preprocessing performed with Quantile Transformation, seen in Figure 3.

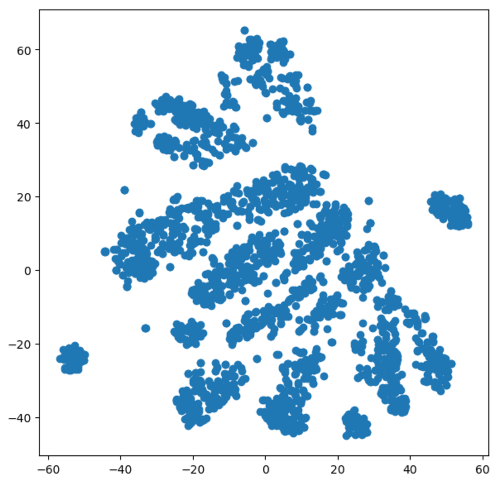

Figure 1: Visual representation of data preprocessed using Standard Scaler.

In Figure 1, we can see well forming clusters, that appear to spread out more in a pedal shape. This is not a necessarily poor way to cluster the data, however this method was not as effective as clustering as some of the others.

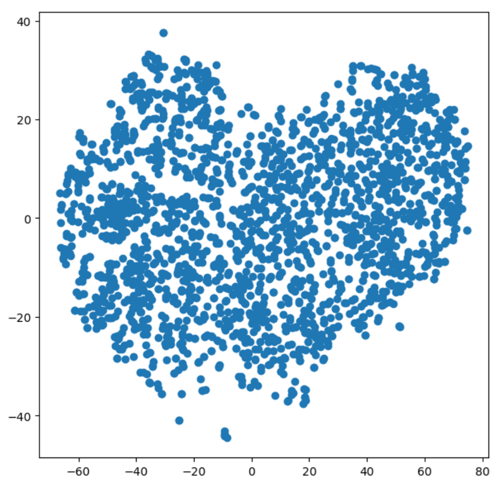

Figure 2: Visual representation of data preprocessed using Robust Scaler.

Figure 2 contains the cluster for the worst performing preprocessing method, creating seemingly one complete cluster rather than the other preprocessing methods which created smaller more effective clusters.

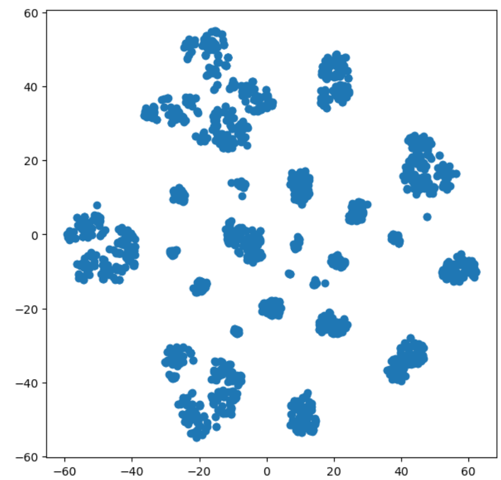

Figure 3: Visual representation of data preprocessed using Quantile Transformation.

Figure 3 created what appears to be the most effective method at preprocessing for this dataset, Quantile Transfomation. Creating much smaller and more compact cluster compared to the other methods.

Lastly, Log Transformation, performed similarly to how the Standard Scaler preprocessing performed. It created smaller clusters within a larger cluster of data.

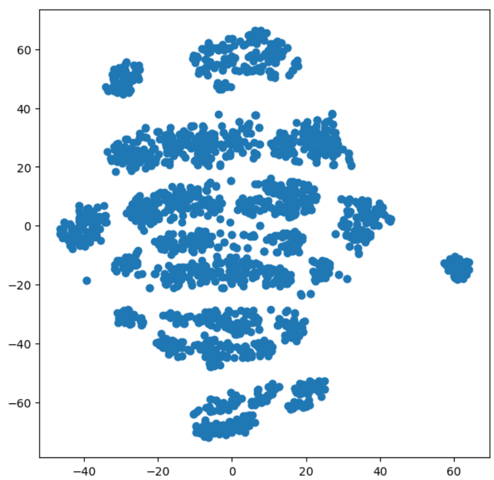

Figure 4: Visual representation of data preprocessed using Log Transformation.

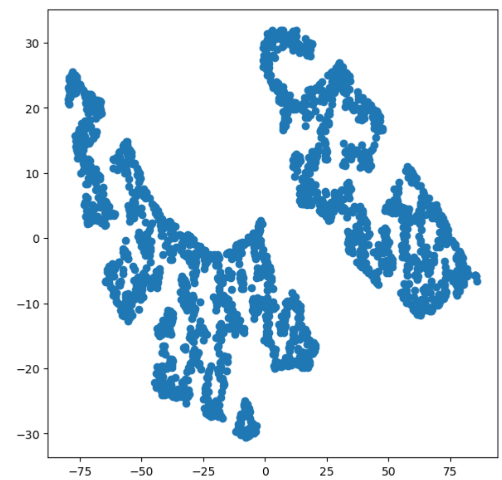

Figure 5: Visual representation of data with t-SNE feature reduction with 2 n_clusters.

2.2 Feature Reduction

Feature reduction was a necessary addition to this project, with the preprocessed dataset contain 33 features. Two feature reduction were tested to determine which method would be most effective. PCA was the first method tested and the results were very poor. Which led to the discovery of the T-distributed Stochastic Neighbor Embedding (t-SNE) method. This method turned out to be far more effective for this dataset. Figure 5 depicts the t-SNE feature reduction with the n_clusters parameter set to 2. This was the most effective feature reduction method. Additional n_cluster values were tested, n_clusers set to 3 is shown in Figure 6. We can see that the clustering was far less effective using this method.

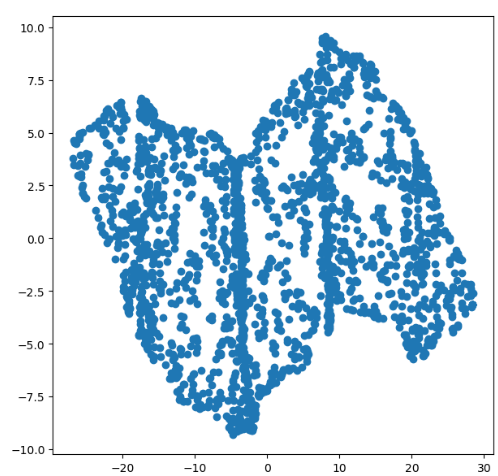

Figure 6: Visual representation of data with t-SNE feature reduction with 3 n_clusters.

2.3 Clustering

2.3.1 K-Means

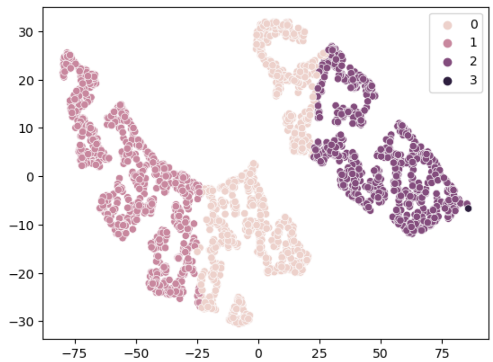

Figure 7: Visual representation of dataset clustered with K-Means method with t-SNE feature reduction applied.

2.3.2 Hierarchical

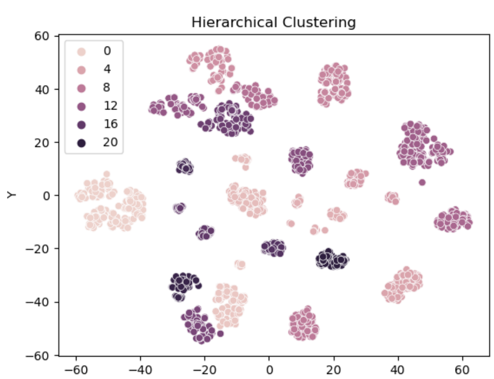

Figure 8: Visual representation of dataset clustered with Hierarchical method with t-SNE feature reduction and Quantile Transformation applied.

2.3.3 DBSCAN

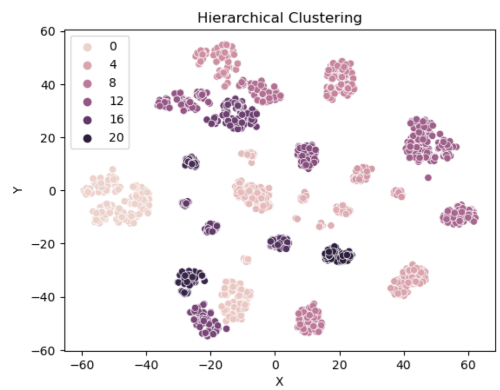

Figure 9: Visual representation of dataset clustered with DBSCAN method with t-SNE feature reduction and Quantile Transformation applied.

3 Results

4 Conclusion

References

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

[2] J. Zhao, "Customer segmentation application based on k-means," *Applied and Computational Engineering*, vol. 47, pp. 242–247, Mar. 2024. DOI: `10.54254/2755-2721/47/20241400`.

[3] O. Sokol and V. Holý, "The role of shopping mission in retail customer segmentation," *International Journal of Market Research*, vol. 63, no. 4, pp. 454–470, 2020. DOI: `10.1177/1470785320921011`.

[4] J. John, O. Shobayo, and B. Ogunleye, "An exploration of clustering algorithms for customer segmentation in the uk retail market," *Analytics*, vol. 2, pp. 809–823, Oct. 2023. DOI: `10.3390/analytics2040042`.

[5] N. Hicham and S. Karim, "Analysis of unsupervised machine learning techniques for an efficient customer segmentation using clustering ensemble and spectral clustering," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 10, 2022. DOI: `10.14569/ijacsa.2022.0131016`.

[6] B. Turkmen, "Customer segmentation with machine learning for online retail industry," *The European Journal of Social and Behavioural Sciences*, vol. 31, no. 2, pp. 111–136, 2022. DOI: `10.15405/ejsbs.316`.