Customer segmentation using their spending personality

Final Report: DS5230 Unsupervised Machine Learning

Authors: Reem Ghabayen, Nidhi Bodar , Foram Shah

Abstract

Customer segmentation using spending personality helps an organization/company for the betterment of sales and production. In this project, we have used Kaggle uncleaned dataset, after cleaning it to create a meaningful conclusion. Through this dataset, we seek the best way to segment customers as per their spending on particular projects by their age, education level, marital status, etc. We have used various types of clustering such as K-means, GMM, and fast clustering to know which algorithm works better reducing the computational cost. We have performed Exploratory data analysis before performing modeling to have an understanding of how models can capture various patterns and compare the results which are evident by humans. Although, after performing modeling we received detailed results of how each category has an impact on the other. As well as what kind of combinations of customers can yield fruitful results. This information can be useful to merchants who are willing to increase their profit by knowing their customers well.

1 Introduction

In any organization, it is very important to know who their targeted customers are to maintain the quantity and types of products. Once it is known, it can be beneficial for the organization as they can think about what other products their targeted customers would tend to buy. For example, if our targeted customers are unmarried Graduate students then we should be focusing more on groceries like Milk, eggs, bread, etc and if our targeted customers are Parents with young kids then we can increase the number of Chocolates, candies, fruits, etc in the store. Customer segmentation helps us profile the customers and it can be known what category of customers tend to spend more on what items hence such items can be increased or placed differently inside the store. We have used Clustering with the help of different clustering algorithms such as K-means, and GMM applied Fast clustering techniques and derived which one works better for our dataset. The dataset consists of many features describing customers which can be used to categorize them with the help of clustering such as a customer's marital status, education level, age, number of kids at home, income, spending history about particular products, and so on. After clustering, we calculated Silhouette scores, Davies-Bouldin Index and Calsinki Index to understand which clustering algorithms should be used to get the best results.

2 Background

The customer classification clustering modeled as an unsupervised clustering problem followed the same set of steps involved in building machine learning models, which are: are:

- Data Collection
- Data Prepossessing
- Model Selection
- Training the model
- Evaluate Performance
- Profiling the clusters

For convenience, we have combined some of these steps in this report to better understand the process.

2.1 Preprocessing

The data from Kaggle had about 29 features, a few of which were unnecessary for our project. We initially checked which features had null or empty values and removed those features as they might create anomalies later on. The dataset contains the features such as year of birth, income, marital status, kids at home, recency which means the number of days before which they made their last purchase, and date of Customer meaning the date when the customer was enrolled with the company. We also have columns stating the amount a customer spends on products such as fish, gold, meat, sweets, etc. Features like the number of web purchases, number of store purchases, and number of catalog purchases were removed for simplicity as our main focus was on how much a particular customer would spend on any product regardless of the medium of purchase. The following figure refers to the columns initially present in the dataset. fig-1

```
df.head()
```

	ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines	...	NumWebVisitsMonth	AcceptedCmp3	Acc
	5524	1957	Graduation	Single	58138.0	0	0	04-09-2012	58	635	...	7	0	
	2174	1954	Graduation	Single	46344.0	1	1	08-03-2014	38	11	...	5	0	
	4141	1965	Graduation	Together	71613.0	0	0	21-08-2013	26	426	...	4	0	
	6182	1984	Graduation	Together	26646.0	1	0	10-02-2014	26	11	...	6	0	
	5324	1981	PhD	Married	58293.0	1	0	19-01-2014	94	173	...	5	0	

ows × 29 columns

Figure 1: Original Dataset

Then we made some box plots to identify if

we have any outliers. We found out that there were some customers whose age was more than 100 and certain customers having their income of more than 600000 which we avoided because it might be problematic while using clustering algorithms like Kmeans. fig-2.

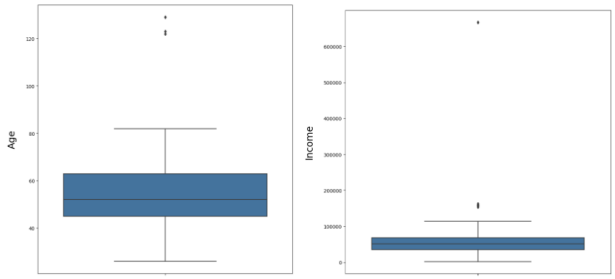
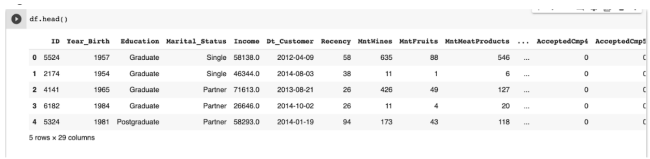


Figure 2: Age and Income Outlier

We calculated the age of customers by subtracting the birth year from 2022 in order to get the current age and later on converted the age into age groups as it would be easier for us to distinguish them with the help of age groups. Further on the customers were categorized as old customers and new customers based on the customer data. Customers till 2012 were considered old customers and the ones after 2012 i.e from 2013 were considered as new customers. This can help us in the future to understand whether the old customers are still interested in our products or what are those products which can attract our new customers. Later on, there were different features for having a teen or a kid which we combined and changed into the number of children in the house. For marital status, we had 8 different values which we changed to Single and Partner as it would be much easier to categorize them this way. There were also different education levels which were then converted into 3 categories as Graduate, Post Graduate, and Undergraduate. After the above steps, the dataset finally looks as follows as in fig-3.



df.head()

	ID	Year_Birth	Education	Marital_Status	Income	Dt_Customer	Recency	MntWines	MntFruits	MntMeatProducts	...	AcceptedCmp4	AcceptedCmp5
0	5524	1957	Graduate	Single	58138.0	2012-04-09	58	635	88	546	...	0	0
1	2174	1954	Graduate	Single	46344.0	2014-08-03	38	11	1	6	...	0	0
2	4141	1965	Graduate	Partner	71613.0	2013-08-21	26	426	49	127	...	0	0
3	6182	1984	Graduate	Partner	26646.0	2014-10-02	26	11	4	20	...	0	0
4	5324	1981	Postgraduate	Partner	58293.0	2014-01-19	94	173	43	118	...	0	0

5 rows × 29 columns

Figure 3: Final Dataset

2.2 Exploratory Data Analysis

After taking essential steps towards preprocessing, we performed EDA to get a more clear understanding of our data and to know how it is distributed. We plotted a graph for each category to know the frequency of unique categories in one column in the dataset. There are a few categories on which we perform feature engineering to ensure we receive results that can be exploited by organizations such as Martial status, Education, and Kids. To gain insights on what proportions of age customers we have, we created groups of ages and plotted the frequency of that as well. By combining all the separate expenses such as fruits, wine, meat, fish, and gold products, we computed one column named Expenses and plotted various graphs for different categories vs Expenses. This gave us an idea of how particular categories such as education level or having kids affect expenses on what level. Also, we plotted graphs of age vs each expense category to know what group of people are spending most on individual products. Moreover, we plotted an income graph for the age group of people to receive information about their average income of theirs. Looking at the fig-4 below.

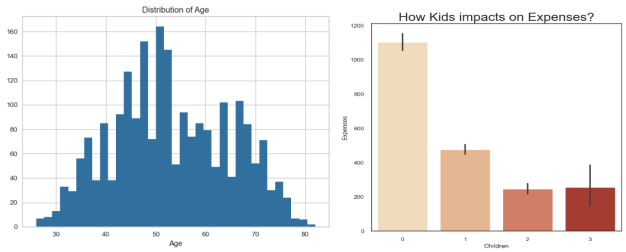


Figure 4: Exploratory Data Analysis Figures

As we can see from the first above image, the histogram shows how age has been distributed across our data. Also, the second graph depicts how having kids is impacting the expenses of customers. Similarly, we have computed graphs for each and every category to visualize it in a proper way.

3 Project description

After prepossessing the data was done, the data was in a tidy format to start modeling. Before modeling, we ended up with 33 variables so we wanted to reduce dimensionality using principal component analysis (PCA). When applying PCA Using Sklearn Library, we saw that 95 percent of the variance was explained using three principal components as demonstrated in fig-5.

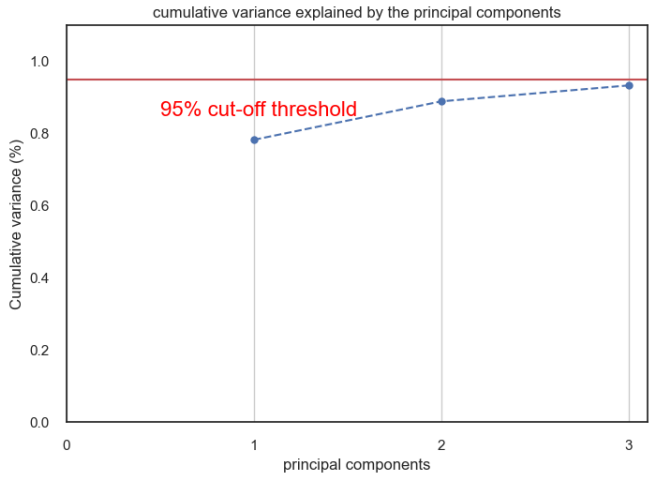


Figure 5: Cumulative Variance Explained by the Principal Components

Another step that was taken previously to the cluster using K-Means was using the Elbow Method to optimize the number of clusters used. The Elbow Method uses K−means by fitting the model with a variety of values, the KElbowVisualizer applies the "elbow" method to assist in choosing the optimal number of clusters, for us the optimal was four. The "elbow" (the point of inflection on the curve) is a solid sign that the underlying model fits well at that point if the line chart resembles an arm in fig-6. below. For applying the Elbow method we used both makeblobs and KMeans from sklearn.

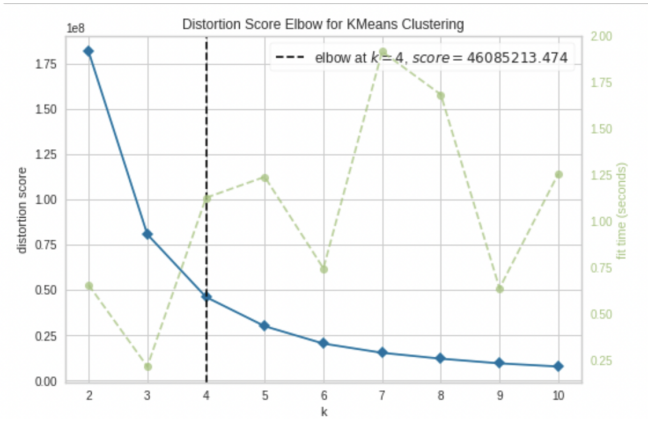


Figure 6: Cumulative Variance Explained by the Principal Components

Using Sklearn we implemented Kmeans clustering, Hierarchical agglomerative clustering, Dbscan, and Gaussian Mixture Models. For each model, we plotted the 3d clusters using seabornFor each of the mentioned models we also calculated and compared their Silhouette, Calinski, and Davies Bouldin scores to measure the performance of the clustering models Demonstrated in the results.

The model that performed the best scores was the Kmeans. We choose K Means to implement the Fast clustering Techniques. We can see that when using clustering algorithms the computational cost can be high that's where fast clustering comes to play, Fast clustering is a technique that reduces the computational cost where the cost is proportional to the size of the sample batch, so we do not iterate over the entire dataset at once. Each batch of data is collected on each iteration to update the clusters. With this technique, we reduce the computational cost, where the cost is proportional to the size of the sample batch, at the cost of lowering the performance of the clustering algorithm. We implemented fast clustering using K Means where we built the fit and predict models from scratch and tested out euclidean and cosine similarity metrics.

4 Empirical Results:

After the examination of data with exploratory data analysis before modeling, we received a basic understanding of how data was distributed. The frequency of people in relationships is more than people who are single. A lot of them have only one kid. Most of them fall between the ages of 45 to 55. Overall, a group of people between having ages of 75 to 85 consumes all the products more than other groups of people. However, all ages customers have a nearly similar pattern in purchasing gold. It can be possible because customers see gold as an investment. One other finding is that customers who are single spend the most of their money which can be corroborated by the expenses of having kids since having 0 kids result in more spending. People who are doing their Ph.D. or Master's have high expenses as well.

As we can notice from the graphs below fig-7, the group of people ages 75 to 85 has higher income and high expenses in all the categories mentioned in our dataset which is kind of intuitive to corroborate two graphs, income vs age group and expenses vs age group. Similarly, people aged between 35 to 45 have comparatively low incomes and expenses. Therefore, it is clearly visible that these two columns have an impact on each other.

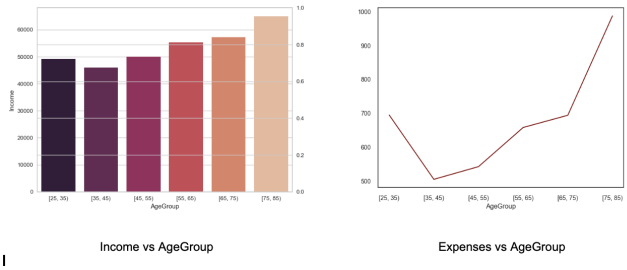


Figure 7: Income and Expense Vs Age Group

Looking at the Table in fig-8 below we can see the silhouette, the calinski, and the Davies score for each of the algorithms we used, plus the scores for the fast clustering technique when applied to Kmeans.

As we can see the silhouette scores were about the same, but the closer to +1 they are the better, so we can definitely see that Kmeans had the highest followed by KMeans Fast clustering. For Calinski again the higher the better the separation of clusters, again we see that Kmeans followed by Fast clustering Kmeans. On the other hand for davies Bouldin the lower the score the better, we can see that the lowest was the Fast clustering Kmeans, followed by regular Kmeans.

Model	silhouette_score	calinski_harabasz_score	davies_bouldin_score
K Means	0.645857697911408	12088.1857312	0.49584790377
Hierarchical Agglomerative	0.608433090329519	8457.51392442	0.51250250840
GMM	0.478505340984214	5914.96820314	0.63340280767
Fast Clustering KMeans	0.645482636409098	12072.0390685	0.49512533576

Figure 8: Evaluation Metrics

When Running the DBScan algorithms, it did not produce separate clusters, so we could not get results for the different scores, which might be because DBScan doesn't work well over clusters with different densities, and it needs a better selection of its parameters. For the future would to produce clusters for DBscan we would use grid search for hyperparameter tuning summary DBScan is not able to cluster this properly because it is not able to cluster data with large variances in density, data with non-spherical shapes, or data with overlapping clusters, which in our case 2 of them hold.

We Can Also see the Gmm had the second lowest performance from the different scores, it can be because of GMMs making soft partitions, also GMMs perform weaker when categorical variables are present, and needs a sufficient amount of data for each cluster.

Evaluating the Clusters:

Before the clustering profiling, we evaluated some variables against the clusters to see each variable's density in each cluster. Below are some of the graphs with interesting findings.

As we can see in the example below in fig-9 , cluster 1 had the highest expenses, meaning the highest spending customers, it also demonstrated the highest customers with higher income.

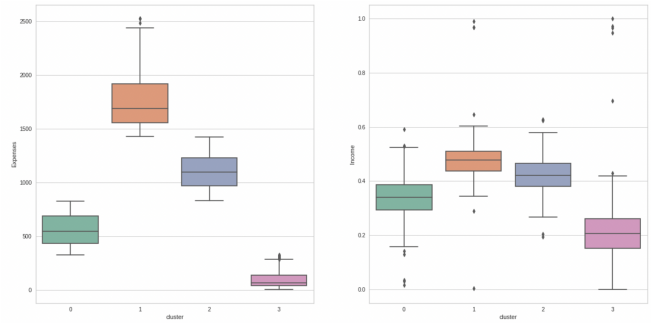


Figure 9: Income and Expense Vs Age Group

Cluster Profiling:

is done by using the input variables you utilized for the cluster analysis, profiling entails creating descriptions of the clusters. We were able to profile our clusters as followers. For a number of selected variables like Children, Age, MaritalStatus-Partner, MaritalStatusSingle, EducationGraduate, EducationUndergraduate, CustomertimespentNew-Customer, and CustomertimespentOldCustomer.

We visualized the densities of each variable in each cluster and came up with the following profiling:

Cluster 0:(Lowest Expenses In general, density Very High, lowest Income) A parent with more kids at home(teens and children, In the older Age group, the Majority have a partner, Above Undergraduates Mostly old Customers

Cluster 1:(Highest Expenses in general, Density Very sparse, Highest Income)) One to Zero Kids at Home, average age group, mostly have a partner, No undergraduates, Majority Old Customers

Cluster 2:(Second Highest Spending Group) One to Zero Kids, Average Age group, Mostly have a partner, Mostly Graduate, Zero Undergraduates, Average Customers

Cluster 3:(Second to last spending group) The majority 1 Kid at Home, Average age group, Mostly Have a partner, half have a graduate degree half with no degree, Average Customers

After Profiling, in addition, we created a table that contained all Variable Percentage Allocation Per Cluster. Demonstrated below is an example fig-10

index	cluster	variable	Mean_Column	percentage
0	0	Year_Birth	1970.9083175803403	47.878473662856926
1	0	Income	0.21088261513972417	32.2637700751864
2	0	Recency	0.4950831567088657	47.82396178214718
3	0	MntWines	0.030379958394372224	7.106217125527177
4	0	MntFruits	0.025044884156130365	9.05375937913154
5	0	MntMeatProducts	0.013596887756499821	6.716377513758925
6	0	MntFishProducts	0.02748319477968117	9.043096098656324
7	0	MntSweetProducts	0.020512561508824082	9.504070068698808
8	0	MntGoldProds	0.04531267482877821	15.838170515828908
9	0	NumDealsPurchases	2.11436672967863	43.504472967716836

Figure 10: Snippet of Variable Cluster Allocation

Related Work:

As there exists multiple research and theories on customer segmentation, we went through research papers that explained clustering on similar kinds of data as we have. After a thorough examination of all the clustering algorithms, we noticed that some of

them used Spectral Clustering to differentiate customers. Although, their conclusion was similar such as this clustering algorithm was working relatively slowly compared to others. Also, the results which were obtained by performing the Spectral clustering weren't easy to explain. Therefore, we decided to drop spectral clustering for our project.

5 Conclusion:

In conclusion, unsupervised clustering in this project was used. The elbow technique, Kmeans, DBSCAN, GMM, and hierarchical clustering were applied to determine the ideal number of clusters utilizing PCA's dimensionality reduction. The Kmeans Model variation to develop a quicker clustering method was selected that allowed us to train huge datasets in exponentially less time with little accuracy loss. Further consumer profiling and EDA allowed us to determine their characteristics, such as expenses, incomes, and family groups, which helped the marketing organizations further divide the customers into groups and target the correct audience. All in all, Looking at our different cluster profiles for our project our target customers would be from cluster one since they demonstrated the highest income and spending customers. For future work, we can build an application on one of the frameworks where the owner can add the products to the application and their customer details will be stored every time they purchase anything. With the help of this application, the owner can predict how much that customer will spend on what products next time depending on their age group, education level, and other such details about the customer.

6 Github Link and Statment of Contribution:

The code for this project is hosted on: `https://github.com/ReemGhabayen/Customer_Segmentation_USML`

The project was divided among three group members equally and was completed and curated by everyone. As per the general project requirements, data cleaning and preprocessing have been performed by Foram Shah. Exploratory Data Analysis and Feature Engineering have been performed by Nidhi Bodar. Reem Ghabayen has contributed to Cluster Modeling and Cluster profiling.

References

[1] Clustering in Recurrent Neural Networks for Micro-Segmentation using Spending Personality. (2021, December 5). IEEE Conference Publication — IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/9659905

[2] Wu, R.-S., Chou, P.-H. (2010, November 14). Customer segmentation of multiple category data in e-commerce using a soft-clustering approach. Electronic Commerce Research and Applications. Retrieved October 24, 2022, from https://www.sciencedirect.com/science/article/pii/S1567422310000888

[3] Roman, V. (2021, December 9). Unsupervised Learning Project: Creating Customer Segments. Medium. https://towardsdatascience.com/unsupervised-learning-project-creating-customer-segments17c4b4bbf925

[4] Ateşli, H. (2022, January 5). Customer Segmentation for Financial Services - Analytics Vidhya. Medium. https://medium.com/analytics-vidhya/customer-segmentation-for-financial-services-58fbfc417669

[5] NCBI - WWW Error Blocked Diagnostic. (n.d.). Retrieved October 24, 2022, from . https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6820452/

[6] Nurma Sari, Juni and Nugroho, Lukito and Ferdiana, Ridi and Santosa, Paulus. (2016). Review on Customer Segmentation Technique on E-commerce. Advanced Science Letters. 22. 3018-3022. 10.1166/asl.2016.7985.

[7] Cooil, Bruce and Aksoy, Lerzan and Keiningham, Timothy. (2007). Approaches to Customer Segmentation. Journal of Relationship Marketing. 6. 9-39. 10.1300/J366v06n0302

[8] Jun Wu, Li Shi, Wen-Pin Lin, Sang-Bing Tsai, Yuanyuan Li, Liping Yang, Guangshu Xu, "An Empirical Study on Customer Segmentation by Purchase Behaviors Using a RFM Model and K-Means Algorithm", Mathematical Problems in Engineering, vol. 2020, Article ID 8884227, 7 pages, 2020. https://doi.org/10.1155/2020/8884227