Unsupervised Learning Assignment - Customer Personality Analysis

Overview - Main objective

We are working for a company that owns and manages a grocery store and we are part of its data analytics division that supports various functions such as marketing, purchasing and warehousing. The head of the analytics department has assigned us a project from the marketing team, to identify different customer personality clusters, in other words to segment clients in terms of their purchases and shopping habits. The goal of this analysis is to help the company better understand its customers and allow the marketing team to better promote products to specific client types. It also makes it easier for the business to understand the needs, behaviours and concerns of the respective customer segments and more easily modify the product line.

For this analysis, we will leverage various unsupervised learning techniques that will allow us to cluster the customers in groups. The aim is to find which model will produce reasonable segments that the marketing team will be able to act upon, by creating more personalized suggestions. The main objective of this report is to provide an overview of the data that was used for this analysis, go through the steps taken to clean, analyze and understand the data, the models that were trained and how these were compared against each other as well as our final recommendation towards the head of analytics and the marketing team. The rest of the report is structured as follows: the next two sections describe a) the data used for this analysis and b) the steps taken to clean and analyze the data. This will allow us to anticipate potential issues with our data and treat them accordingly so that we can prepare it for our clustering algorithms. Potential hypotheses about the data will be presented and their validity will be reviewed in the following sections. Then we will provide details on the different clustering algorithms that were used in our analysis and compare their benefits and shortcomings. Finally, our findings are presented and potential flaws are recognized, including suggestions for future research.

Data

For this analysis, we leverage data on customer personality that can be found here. There are entries about 2240 customers, each one identified by a unique ID in the dataset. For each customer, a variety of data is available, spread across 28 columns. Information includes year of birth, number of kids at home (and at what relative age, e.g. kids, teens, etc.), gender, marital status, level of education, income and various metrics on purchase habits such as amount of meat/fruits/vegetables/etc. bought. See below a complete data dictionary:

- **People**
  ID: Customer's unique identifier
  Year_Birth: Customer's birth year
  Education: Customer's education level
  Marital_Status: Customer's marital status
  Income: Customer's yearly household income
  Kidhome: Number of children in customer's household
  Teenhome: Number of teenagers in customer's household
  Dt_Customer: Date of customer's enrollment with the company
  Recency: Number of days since customer's last purchase
  Complain: 1 if the customer complained in the last 2 years, 0 otherwise
- **Products**
  MntWines: Amount spent on wine in last 2 years
  MntFruits: Amount spent on fruits in last 2 years
  MntMeatProducts: Amount spent on meat in last 2 years
  MntFishProducts: Amount spent on fish in last 2 years
  MntSweetProducts: Amount spent on sweets in last 2 years
  MntGoldProds: Amount spent on gold in last 2 years
- **Promotion**
  NumDealsPurchases: Number of purchases made with a discount
  AcceptedCmp1: 1 if customer accepted the offer in the 1st campaign, 0 otherwise
  AcceptedCmp2: 1 if customer accepted the offer in the 2nd campaign, 0 otherwise
  AcceptedCmp3: 1 if customer accepted the offer in the 3rd campaign, 0 otherwise
  AcceptedCmp4: 1 if customer accepted the offer in the 4th campaign, 0 otherwise
  AcceptedCmp5: 1 if customer accepted the offer in the 5th campaign, 0 otherwise
  Response: 1 if customer accepted the offer in the last campaign, 0 otherwise
- **Place**
  NumWebPurchases: Number of purchases made through the company's website
  NumCatalogPurchases: Number of purchases made using a catalogue
  NumStorePurchases: Number of purchases made directly in stores
  NumWebVisitsMonth: Number of visits to company's website in the last month

We note that the data is almost fully complete. For 24 customers income information is not available, representing around 1% of the overall dataset. Summary statistics for all variables can be seen below, while distribution details will be covered in the following section:

People

statistic str	ID f64	Year_Birth f64	Education str	Marital_Status str	Income f64
"count"	2240.0	2240.0	"2240"	"2240"	2216.0
"null_count"	0.0	0.0	"0"	"0"	24.0
"mean"	5592.159821	1968.805804	null	null	52247.251354
"std"	3246.662198	11.984069	null	null	25173.076661
"min"	0.0	1893.0	"2n Cycle"	"Absurd"	1730.0
"25%"	2829.0	1959.0	null	null	35322.0
"50%"	5462.0	1970.0	null	null	51390.0
"75%"	8427.0	1977.0	null	null	68487.0
"max"	11191.0	1996.0	"PhD"	"YOLO"	666666.0

statistic str	Kidhome f64	Teenhome f64	Dt_Customer str	Recency f64	Complain f64
"count"	2240.0	2240.0	"2240"	2240.0	2240.0
"null_count"	0.0	0.0	"0"	0.0	0.0
"mean"	0.444196	0.50625	null	49.109375	0.009375
"std"	0.538398	0.544538	null	28.962453	0.096391
"min"	0.0	0.0	"01-01-2013"	0.0	0.0
"25%"	0.0	0.0	null	24.0	0.0
"50%"	0.0	0.0	null	49.0	0.0
"75%"	1.0	1.0	null	74.0	0.0
"max"	2.0	2.0	"31-12-2013"	99.0	1.0

Products

statistic str	MntWines f64	MntFruits f64	MntMeatProducts f64	MntFishProducts f64
"count"	2240.0	2240.0	2240.0	2240.0
"null_count"	0.0	0.0	0.0	0.0
"mean"	303.935714	26.302232	166.95	37.525446
"std"	336.597393	39.773434	225.715373	54.628979
"min"	0.0	0.0	0.0	0.0
"25%"	24.0	1.0	16.0	3.0

statistic str	MntWines f64	MntFruits f64	MntMeatProducts f64	MntFishProducts f64
”50%”	174.0	8.0	67.0	12.0
”75%”	504.0	33.0	232.0	50.0
”max”	1493.0	199.0	1725.0	259.0

statistic str	MntSweetProducts f64	MntGoldProds f64
”count”	2240.0	2240.0
”null_count”	0.0	0.0
”mean”	27.062946	44.021875
”std”	41.280498	52.167439
”min”	0.0	0.0
”25%”	1.0	9.0
”50%”	8.0	24.0
”75%”	33.0	56.0
”max”	263.0	362.0

Promotion

statistic str	NumDealsPurchases f64	AcceptedCmp1 f64	AcceptedCmp2 f64	AcceptedCmp3 f64	AcceptedCmp4 f64
”count”	2240.0	2240.0	2240.0	2240.0	2240.0
”null_count”	0.0	0.0	0.0	0.0	0.0
”mean”	2.325	0.064286	0.013393	0.072768	0.074554
”std”	1.932238	0.245316	0.114976	0.259813	0.262728
”min”	0.0	0.0	0.0	0.0	0.0
”25%”	1.0	0.0	0.0	0.0	0.0
”50%”	2.0	0.0	0.0	0.0	0.0
”75%”	3.0	0.0	0.0	0.0	0.0
”max”	15.0	1.0	1.0	1.0	1.0

statistic str	AcceptedCmp5 f64	Response f64
”count”	2240.0	2240.0
”null_count”	0.0	0.0
”mean”	0.072768	0.149107

statistic str	AcceptedCmp5 f64	Response f64
”std”	0.259813	0.356274
”min”	0.0	0.0
”25%”	0.0	0.0
”50%”	0.0	0.0
”75%”	0.0	0.0
”max”	1.0	1.0

Place

statistic str	NumWebPurchases f64	NumCatalogPurchases f64	NumStorePurchases f64	NumWebVisitsMonth f64
”count”	2240.0	2240.0	2240.0	2240.0
”null_count”	0.0	0.0	0.0	0.0
”mean”	4.084821	2.662054	5.790179	5.316518
”std”	2.778714	2.923101	3.250958	2.426645
”min”	0.0	0.0	0.0	0.0
”25%”	2.0	0.0	3.0	3.0
”50%”	4.0	2.0	5.0	6.0
”75%”	6.0	4.0	8.0	7.0
”max”	27.0	28.0	13.0	20.0

Data Exploration and Data Cleaning

The high level view of our data already highlighted potential issues that we need to look into and remedy before proceeding with our modeling approach (e.g.)

- Brief summary of data exploration and actions taken for data cleaning or feature engineering. **Pending**

Unsupervised Learning Models - Clustering Algorithms

Kmeans and faster implementations just for fun, check clusters etc DBScan

- Summary of training at least three variations of the unsupervised model you selected. For example, you can use different clustering techniques or different hyperparameters. **Pending**

- Does the report include a section with variations of Unsupervised Learning models and specifies which one is the model that best suits the main objective(s) of this analysis? **Pending** -Does the report include a section with variations of Unsupervised Learning models and which one is the model that best suits the main objectives of this analysis? Yes, there are at least 3 different models. One of them is presented as the better alternative, and some findings are presented.

Findings

- Does the report include a clear and well presented section with key findings related to the main objective(s) of the analysis? **Pending**
- A paragraph explaining which of your Unsupervised Learning models you recommend as a final model that best fits your needs in terms. **Pending**
- Summary Key Findings and Insights, which walks your reader through the main findings of your modeling exercise. **Pending**
- Does the report include a clear and well presented section with key findings about the problem and next steps?

Yes. Takeaways and findings derived from the model are well presented. The quality of insights or the next steps section award this section an extra point.

Flaws and Future Improvements

- Does the report highlight possible flaws in the model and a plan of action to revisit this analysis with additional data or different modeling techniques? **Pending**
- Does the report highlight possible flaws in the model and a plan of action to revisit this analysis with additional data or different predictive modeling techniques?

Yes. There is a comprehensive list of possible flaws of this model and a detailed plan to revisit this with additional data or different predictive modeling techniques. The quality of this section awards it an extra point.

Conclusion

- Summary Key Findings and Insights, which walks your reader through the main findings of your modeling exercise. **Pending**
- Suggestions for next steps in analyzing this data, which may include suggesting revisiting this model or adding specific data features to achieve a better model. **Pending**