Customer Personality Analysis:

Dataset:

Customer Personality Analysis is a detailed analysis of a company's ideal customers. It helps a business to better understand its customers and makes it easier for them to modify products according to the specific needs, behaviours and concerns of different types of customers.

The Customer Personality Analysis dataset [1] by Dr. Omar Romero-Hernandez has been employed for this case study. The dataset consists of about 30 features based on customer-sales data in a supermarket with more than 2500 unique cases to be analysed. The features include personal details of customers like name, birth year, income, number of kids, etc and the amount they spend on different products like fruits, wine, sweets and so on. The list of attributes also includes promotions and campaigns the customer has participated in and the sources from where they did their regular shopping. Here is a list of features present in the dataset.

Attributes:

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

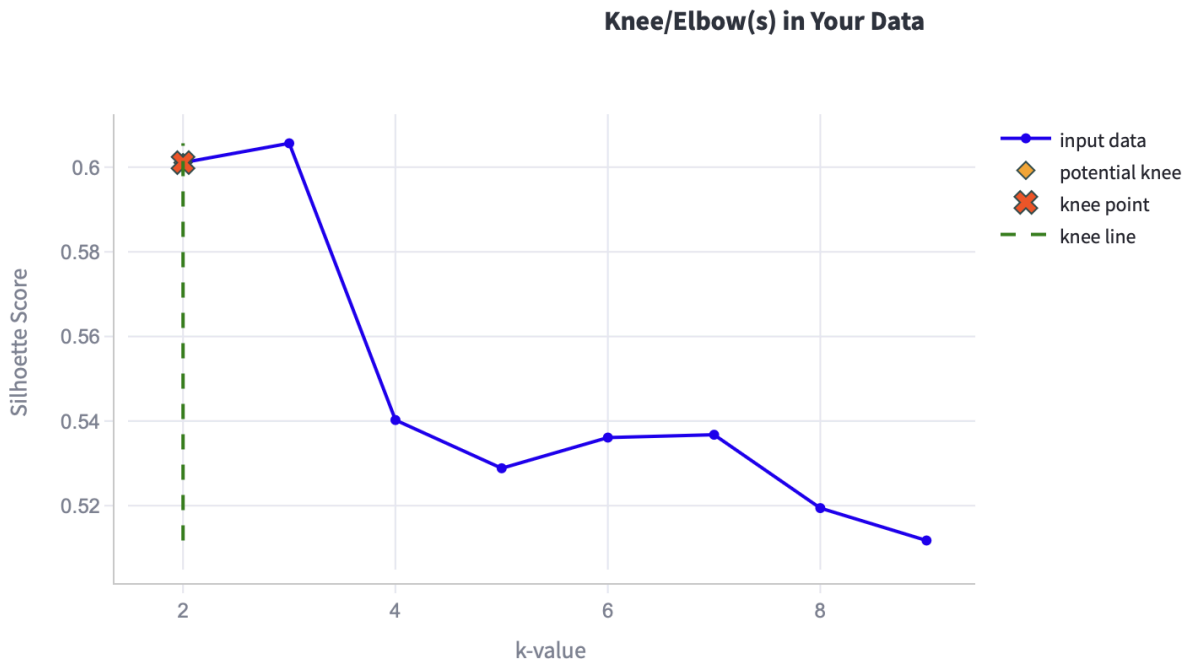
An organisation can alter its product based on the target clients from various customer segments by analysing this data. For instance, a company can determine which customer segment is most likely to purchase the product and then market the product exclusively to that segment rather than investing money to market a new product to every consumer in the company's database.

Data Pre-processing:

Since the data was an open-source one, a thorough data cleaning has to be done. To begin with, all the NaN values were removed and imputed by the median values of that column. The unique features like 'customer ID' were dropped since there was no contribution towards customer segmentation. Features like birth_year and kid_home -teen_home were altered for this specific case in order to get the 'age' and 'number of kids' of the customer. The categorical data were converted into numerical labels for easier processing of the data by the model since it only reads data in a numeric form. However, further modifications in the dataset can be done as required by the client.

Methodology/Workflow

After pre-processing, the attributes are split into either numerical or categorical types based on the threshold that if the number of unique values is <10, then the feature is considered categorical. This thresholding value is user-defined and can be altered according to the dataset. Among all the 29 features, 15 were tagged as numerical features and the rest as categorical in nature. The numerical features are then clustered into 2

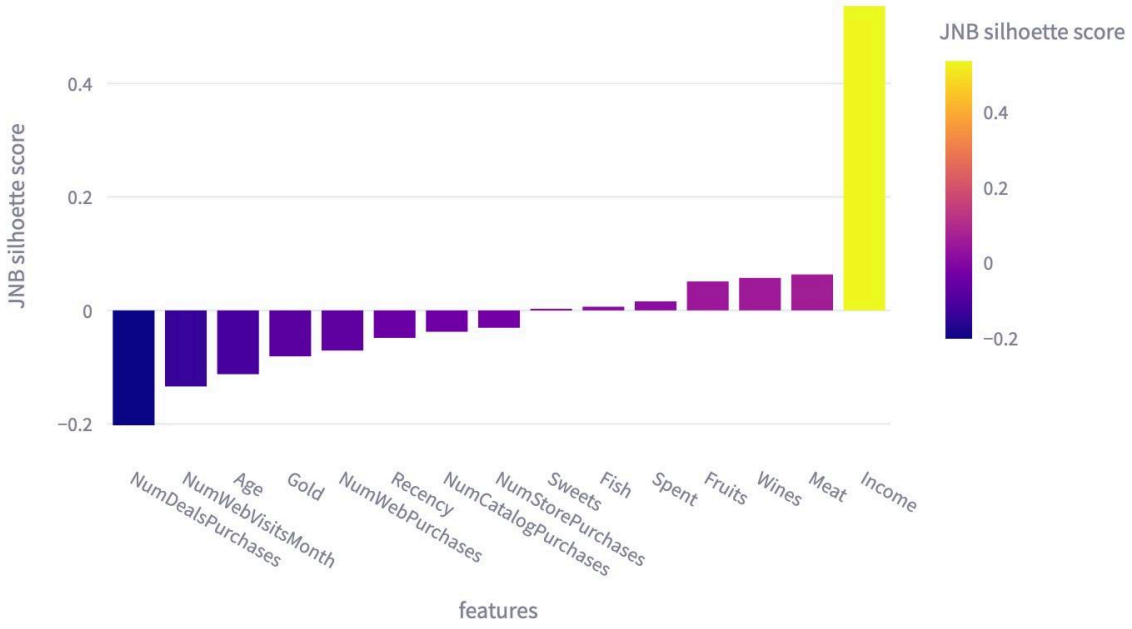


(calculated through the elbow method) groups using JNB clustering[2]. The features are ordered according to their silhouette scores[3] and only the top **30%** are selected for further steps. All possible subsets of these feature spaces are plotted into a feature matrix[4]. For this case study, we considered the combination of "Fruits, Spent" with a silhouette score of 0.694. The entropy of all the categorical features and the selected set is compared.

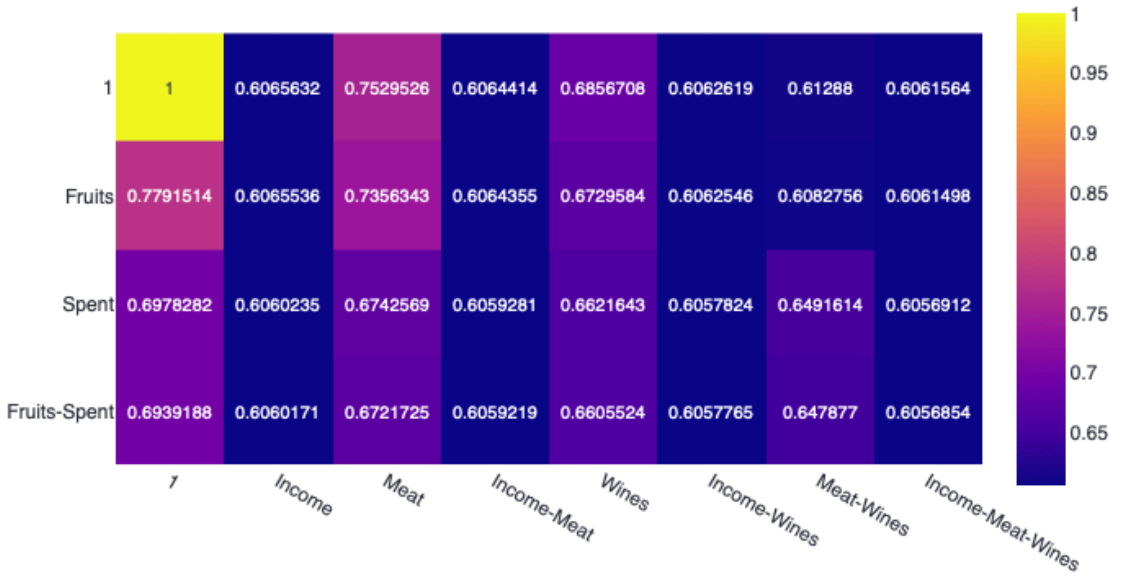
We have used K means clustering to generate all the clusters of the selected numerical features. This is then the joint entropy calculated against all the categorical features. These features are compared according to their entropy in the figure [5]. The maximum entropy is 0.20 for 'children', and the minimum is 0.055 for "AcceptedCmp 5". Depending upon the circumstances, the user can choose one or many categorical features from it. Here, higher values of entropy stand for differences in the representation of features. Lower entropy on the other hand represents similar features to the selected numerical features.

[2]

The Silhouette scores of JNB Clustering are based on the features [3]:



Feature matrix for case study 1 [4]:

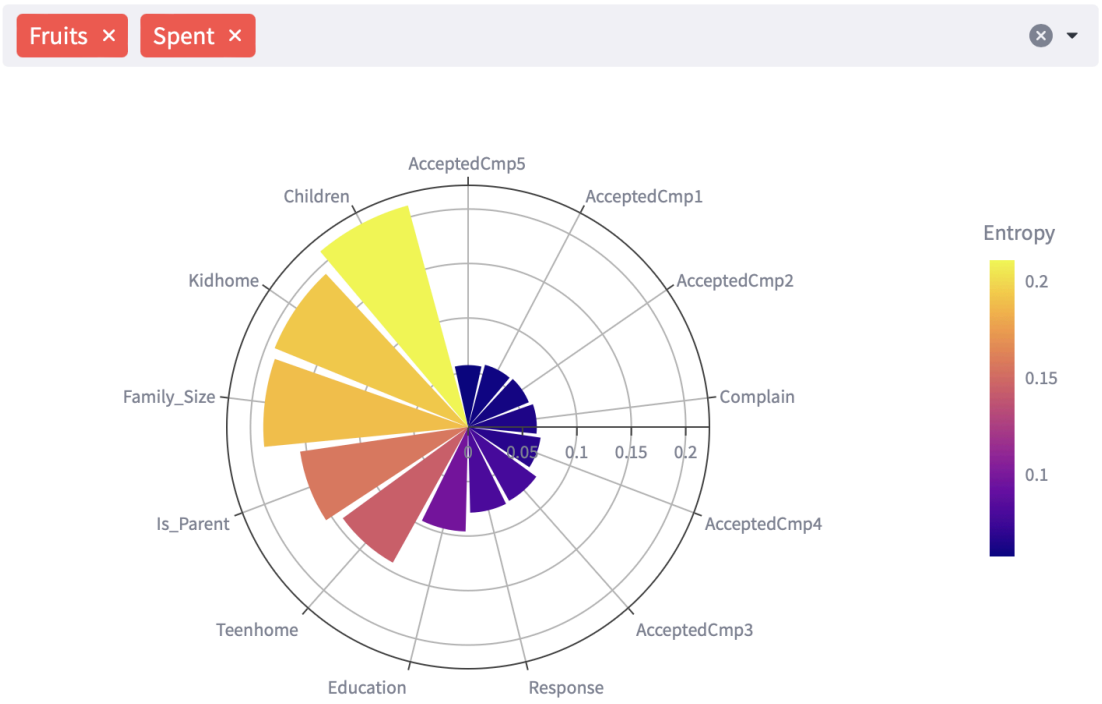


Top performing Attribute sub-sets

	Attribute Subset	Silhoette Score
1	["Fruits"]	0.7792
2	["Meat"]	0.7530
3	["Meat","Fruits"]	0.7356
4	["Spent"]	0.6978
5	["Fruits","Spent"]	0.6939

The entropy of each categorical feature w.r.t. The selected numeric feature/s[5]

Choose the desired features:



Reference Links

1. https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis