CS777 Final Project Report

Kyle Heng

[kheng51@bu.edu](mailto:kheng51@bu.edu)

**Project Background and Research Question**

The focus of this project will be the Customer Personality Analysis dataset publicly available on Kaggle. It is a file named marketing\_campaign.csv that contains data pertaining to different types of customers such as demographic attributes, how much money they spend on different types of purchases, and what platform they make purchases on. Each customer (row) has a unique identifier and 28 other features such as education level, marital status, amount spent on different food categories like fruits or meats, their behaviors towards company promotions, and their number of purchases on different platforms such as online or in store. The data was collected by Dr Omar Romero-Hernandez.The objective of this analysis is to better understand how customers’ demographics align with their purchasing patterns so companies can better tailor their promotions to drive sales. This will be achieved by performing K-Means clustering on the data. By examining the summary statistics of each feature within each cluster, common patterns in spending can be identified and connected to certain demographics.

**Data Format**

The data is a csv file delimited by backslashes. There are 2241 customers(rows) included in the data and 29 atrributes as follows:

* 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

**Data Preprocessing Steps:**

* Dropping Rows with Missing Values
  + Achieved using df.na.drop()
* One-Hot Encoding Categorical Variables
  + Education and Marital Status are the only two categorical variables
  + Pyspark’s StringIndexer fucntion is used to create the index for OneHotEncoder
  + Created pipeline to apply encoding to data
* Converting Rest of Columns to Float
  + Other features were numeric but in string form
* Necessary to move on to feature assembly

**Assembling Features and PCA Dimensionality Reduction**

* Passed cleaned data into Pyspark’s VectorAssembler to combine all features into a single vector for conducting K-Means
* Needed to flatten columns and ensure all data was in numeric form to avoid errors\
* Split data into training and testing sets with an 80-20 split ratio
* Conducted PCA do reduce dimensionality and fit the PCA model to both training and testing sets

**Elbow Method for Determining K**

**A graph with a line

Description automatically generated**

* “elbow” seems to lie around k=7, suggesting that this is the optimal k value for k-means clustering

**K Means Clustering**

* Conducted k-means clustering with k=7 clusters on training data
* Used model to calculate predictions on testing data

**Silhouette Analysis of Clusters**

**A graph of a bar graph

Description automatically generated**

* Clusters 4 and 5 have much lower numbers of points than the other clusters
  + Points may have been inaccurately clustered, or boundaries are unclear
* Focus will be shifted to Clusters 0,1, and 3 since they have higher concentrations of points

**Summary Statistics of Clusters with Higher Silhouettes**

* **Cluster 0**

**A screenshot of a computer

Description automatically generated**

* **Cluster 1**

**A computer screen shot of a number

Description automatically generated**

* **Cluster 3**

**A screenshot of a computer

Description automatically generated**

**Final Insights**

* Customers in Cluster 0 seem to be slightly older and have higher income on average compared to 1 and 3.
  + With the higher income comes an increase in spending overall and more spending towards wine compared to more essential food categories
  + They are also divorced or widowed, potentially influencing the alcohol purchasing as a way of coping
* Customers in Cluster 1 make the least amount of web purchases than the other two clusters with Cluster 3 making the most.
  + People who are older are more likely to make in store purchases than online, likely due to less computer literacy
  + Customers in this cluster are single and spend less overall compared to the other two clusters.
* Customers in Cluster 3 are married and seem to make the highest amount of store purchases and web visits, suggesting that they have the highest overall frequency of spending compared to other marital status groups.