Data Curation Project Proposal

Team Members:

1. Ramitha Kotarkonda - rkotar2@illinois.edu
2. Matthew Guan - mg95@illinois.edu
3. Murali Natarajan - muralin2@illinois.edu

Overview

The goal of this project is to curate the Customer Personality Analysis dataset to analyze a marketing campaign dataset to better understand customer demographics, purchasing behavior, and campaign effectiveness.

Research Objective:

"*What factors drive customer response to marketing campaigns, and how can curated data improve targeted marketing strategies?"*

The use case is to help a retail company optimize marketing efforts by identifying target customer segments for future campaign events. By running targeted campaigns, the retailers can expect a better response to the campaigns and help increase the sale.

Plan

We would like to relate to the USGS Science data lifecycle for planning this project. The key elements of the lifecycle are plan, acquire, process, analyze, preserve, and publication/sharing.

Plan: We are planning to use a dataset called Customer Personality Analysis. It has about 2241 rows where each line represents a customer. This dataset is from Kaggle.com which is a public dataset. The dataset contains the year of birth which could be considered personally identifiable information. However, in this project, year_birth can not be directly linked back to a specific individual because each individual is linked to an ID. For the purpose of this project, we can consider the year of birth as a PII variable and complete the de-identification process. We can store this dataset in GitHub and create reports to share the results.

Acquire: We can acquire the dataset on kaggle.com which is publicly available and is open for analysis. This data set is never expected to get updated. So we won't need to constantly pull the data from the website.

Processing: The processing can be done through python notebooks that will be located in GitHub. We can begin by data cleaning which includes handling for missing or null values, ensuring all the variables are in the right format, and if needed normalize columns. Next, we would need to de-identify the PII variable, year_birth. To de-identify the PII variable, we aim to convert the birth year to age group to make it less identifiable (i.e. 20-29, 30-39 etc). Once the data cleaning process is completed, we would like to convert all categorical variables into numerical variables to have as much data as possible to input into an unsupervised machine learning algorithm.

Analysis: To analyze the data, it would be beneficial to explore relationships between each variable. Once we grasp a basic understanding of the dataset and relationship between each variable. We can create a segmentation model to identify clusters within these customer segments. Since we do not know the different segments the customers can fall into, we would need to use an unsupervised machine learning algorithm. Different algorithms we can use are Kmeans, DBSCAN, and/or HDBSCAN. We can also perform dimensionality reduction such as PCA or UMAP to reduce the number of variables that are being inputted into the model. These different clusters will help us answer our use case which is to identify the common purchase customers.

Since the goal of the project is to determine factors driving customer response to marketing campaigns, we also plan to add a prediction model (Linear Regression, Random Forest) to determine the relevant importance of each input feature to the overall output. This will be done after the clusters have been determined by the unsupervised learning algorithm. We plan to evaluate the performance of various models and compare the results.

Preserve: We would like to store our work in GitHub. This would enable us to have a clear record of changes we have made through commits. And we can easily incorporate a readme for clear documentation of our work so that it is easily reproducible for others.

Publication/Sharing: We would like to create a report to share this information. This report can be also added to GitHub for future users.

Throughout this entire process, we will simultaneously perform the cross-cutting activities. Cross cutting activities in USGS Science data lifecycle include describing, managing quality, and backup and security.

Describing: We are planning to incorporate a readme file within GitHub to provide metadata about the dataset. This will also be a perfect place for data documentation. Furthermore, we will be focusing on providing detailed commits for additional data documentation.

Managing quality: Since we will be using an unsupervised clustering algorithm, we need to ensure that the data is clean and distinct clusters. We can use metrics such as silhouette score to identify the ideal number of clusters based on the data that was imputed.

For the prediction model, we can use mean squared error and R^2 to determine how well the prediction fits the actual data.

Backup and security: We will be using GitHub as our repository for our code. GitHub stores every commit with line-by-line differences between each commit. This will enable us to always revert to the prior commit.

Data Sources

Primary Dataset:

The primary data source for this project is the "*Customer Personality Analysis*" dataset. The dataset includes customer-level information for a marketing campaign, such as demographics, purchasing behavior, and response to campaigns.

Source: Kaggle

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

Dataset Attributes:

Attribute	Description
Customer Data	
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
Product Details	
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
Campaign Details	
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
Channel	
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

Team

Team Members:

1. Ramitha Kotarkonda - rkotar2@illinois.edu
2. Matthew Guan - mg95@illinois.edu
3. Murali Natarajan - muralin2@illinois.edu

Timeline And Responsibilities

S.No	Tasks	Deadline	Responsibility
1	Data Acquisition	Week1	All Team Members
2	Ethical, legal, and policy constraints	Week 2	Matthew Guan
3	Data Cleaning and Transformation	Week2 – Week3	Murali Natarajan
4	Data Governance (De-Identification)	Week3	Murali Natarajan
5	Data Analysis -Unsupervised Machine Algorithm	Week3 – Week 6	Ramitha Kotarkonda
6	Data Analysis - determining feature importance via prediction models	Week4 - Week6	Matthew Guan
7	Progress Report	Week6	All Team Members
8	Managing Data Quality	Week3 - Week6	Ramitha Kotarkonda
9	Artifacts & Workflow	Week6 – Week7	Murali Natarajan
10	Reproducibility & Transparency	Week 7	Murali Natarajan
11	Metadata & Documentation	Week8 – Week9	Ramitha Kotarkonda
12	Final Report	Week 10 - 12	All Team Members

* Week 1 means the first week of the project i.e. (9/15 to 9/19)

Constraints

Privacy: Contains personally identifiable information (names, emails , Year Birth). Must be anonymized before sharing.

Data Quality Issues: Missing values in Income, inconsistent date formats in Dt_Customer.

Limited Data: Dataset may not represent the full customer population.

Ethical/Legal/Policy Constraints: This project must ensure that customer data is handled responsibly and in compliance with relevant privacy laws and ethical guidelines. The collection of data will be done with adherence to intellectual property laws. The analysis of data will be conducted in an ethical manner: avoiding bias when it comes to sensitive variables such as income or household composition.

Gaps

Below are few gaps identified with the Data:

Outdated Dataset: Dataset seems to be outdated and covers a limited time period (2012-2014 based on customer acquisition dates). The dataset does not appear to be frequently updated. Customer trends that existed in the 2010s may not necessarily hold true today.

External Data: The defined usecase would benefit from external data sources, such as market-wide consumer spending trends or Competitor sales analysis.