Data Mining 1

Fall Semester 2024-2025

Project Description

Our Imaginary Enterprise:

WalSmart

WalSmart is a ten-year-old supermarket chain that seeks to provide customers with a diverse range of high-quality products, including groceries, fresh produce, and unique specialty items sourced from around the world. Its mission is to delight customers with well-curated, affordable, and exciting products that might not be available at other large retail stores.

Through aggressive promotions in food magazines, online ads, and local media, WalSmart has built a strong customer base, currently maintaining a database of over 10,000,000 shoppers. Most customers are regular grocery shoppers, but many also seek out specialty and gourmet items. WalSmart sometimes offers a variety of products, ranging from dairy and meat products to wines.

WalSmart collected customer data in the last year but has so far focused on mass marketing. Now, WalSmart wants to "get smart" about its database, start differentiating its customers, and develop more targeted marketing programs.

The company has provided a random sample of ~120,000 of its active customers. These customers have made a purchase during 2023.

Using this data, you will perform a customer segmentation with the goals of:

- Engagement/Value Segmentation: Identify different groups of customers based on how valuable they are to the company (e.g., high spenders, frequent shoppers, etc.). You will need to select relevant variables from the dataset for this analysis.
- Buying Behavior Segmentation: Understand different customer groups based on their product preferences (e.g., fresh produce, gourmet items, household goods, etc.). This analysis will focus on the categories of products that customers tend to buy more frequently.

You can decide if you would like to build one or more different segmentations.

Finally, you will consolidate the findings from both segmentations and propose tailored marketing strategies for each identified customer segment.

Dataset Description

Variables in the ProjectData table

Field	Description
ID_Client	Customer ID
Frequency_items	Quantity of items bought by customer (2023)
Potential_Score	Customer's potential score. Measure of customer's potential for increased purchases (-100 to 100). % from previous year. Computed by other models.
ZIP_Code	Macro Zip Code of customer's address
Longevity_months	Measured from 31st December 2023 backwards
Latitude	Latitude of customer's address
Longitude	Longitude of customer's address
Credit_factor	Customer's credit worth, (0 to 100). Computed by other models.
Relevance_criteria	Relevance to potential upward tier
Store_state	State where the store is located
Recency_in_weeks	Time from last purchase (as of December 31, 2023)
Returns	Monetary value of returns (2023)
ID_Store_last	Point of Sale of last purchase
Gender	Customer's gender
Flagged	Customer's card flagged as risky (potential cloned card)
Promotional_percentage	Percentage of items bought in promotion (value)
Education	Customer's education
Checked_ok	Flag for customer purchases exported from transactional database
Beer	Revenue from sales in the category (2023), in monetary units, net of sales tax
Bottled_Water	Revenue from sales in the category (2023), in monetary units, net of sales tax
Bread	Revenue from sales in the category (2023), in monetary units, net of sales tax
Meat	Revenue from sales in the category (2023), in monetary units, net of sales tax
Dairy	Revenue from sales in the category (2023), in monetary units, net of sales tax
Fresh_Foods	Revenue from sales in the category (2023), in monetary units, net of sales tax
Frozen_Foods	Revenue from sales in the category (2023), in monetary units, net of sales tax
Fruit_Beverages	Revenue from sales in the category (2023), in monetary units, net of sales tax
Pastry	Revenue from sales in the category (2023), in monetary units, net of sales tax
Sodas	Revenue from sales in the category (2023), in monetary units, net of sales tax
Toiletries	Revenue from sales in the category (2023), in monetary units, net of sales tax

Veggies	Revenue from sales in the category (2023), in monetary units, net of sales tax
Wines	Revenue from sales in the category (2023), in monetary units, net of sales tax

Project Deliverables

1. **Written Report** describing:
    a. the results and conclusions you gained from the analysis you performed;
    b. the process by which you arrived at these insights, justifying the decisions you made;
    c. the different clusters you found.

    No need to discuss how the algorithms work (unless you used an algorithm not discussed in the practical sessions).

    The report should be submitted as a pdf file and named using this format where "99" should be replaced with your group number:

    DM1_2324_Group_99_Report.pdf

    Make sure that the report contains the **names** and **student numbers** of all the members of the group.

    Minimum 5 pages, maximum 15 pages of content (excluding cover page, index and appendices).

2. **Colab Notebook** (ipynb file format) containing all the code used to develop your project. Instructions will be provided for how to download your notebook from Google Colab.

    The file should be named using this format, where "99" should be replaced with your group number:

    DM1_2324_Group_99_Notebook.ipynb

    Make sure that the notebook contains a Markdown cell (text cell) with the **names** and **student numbers** of all the members of the group.

3. **Deadline**. 03 January 2024 23:59. A penalty of 10% will be given for each day of delay.

4. **Discussion**. After submitting the projects the students will be called to discuss the project with one of the instructors.