Assignment 3 (Interim Submission)

Business Intelligence

Dante Godolja
Person A
11929150

Lukas Burtscher
Person B
11925939

1 Business Understanding

1.1 Data Source and Scenario

The data source[1] is a dataset from a retail company's marketing campaigns over time. The dataset contains demographic information, purchasing behavior, and campaign response data for a number of customers. A scenario for a business analytics task based on this dataset involves optimizing future marketing campaigns by analyzing customer responses to previous campaigns and identifying patterns in purchasing behavior.

1.2 Business Objectives

The primary business objective is to increase the efficiency and effectiveness of marketing campaigns. The goal is to understand which segments of customers are most responsive to certain types of campaigns and to tailor marketing strategies accordingly to maximize the return on investment (ROI).

1.3 Business Success Criteria

Success will be measured by an increase in the response rate to marketing campaigns and a higher conversion rate, leading to an overall increase in sales revenue. Additionally, a reduction in marketing costs by targeting the right customer segments would also indicate success.

1.4 Data Mining Goals

The data mining goal is to develop a predictive model that can forecast customer response to marketing campaigns. The model should be able to identify the likelihood of customers accepting offers in future campaigns based on their past behavior and demographic characteristics.

1.5 Data Mining Success Criteria

Success in data mining will be evaluated by the accuracy, precision, recall, and F1 score of the predictive model. A successful model will accurately predict campaign acceptance, with a particular focus on identifying true positives while minimizing false positives and false negatives.customers accepting offers in future campaigns based on their past behavior and demographic characteristics.

1.6 AI Risk Considerations

Specific AI risk aspects to consider include ensuring that the model does not perpetuate or amplify biases present in the historical data. The model should be fair and not discriminate against any group of customers. Additionally, explainability of the model's decisions is important for stakeholder trust and for addressing potential ethical concerns.

2 Data Understanding: Data Description

2.1 Attribute Types and Semantics

The dataset includes:

1. Demographic attributes: Age, income, education, marital status, number of children at home.
2. Behavioral attributes: Amount spent on various products, number of purchases made through different channels, recency of last purchase.
3. Campaign attributes: Binary indicators of whether a customer accepted offers in each of the past six marketing campaigns.

2.2 Statistical Properties and Correlations

Statistical analysis reveals right-skewed distributions for spending attributes and income (see **Figure 1**), indicating a

[1] O. Parr-Rud. Business Analytics Using SAS Enterprise Guide and SAS Enterprise Miner. SAS Institute, 2014. https://www.kaggle.com/datasets/rodsaldanha/arketing-campaign

concentration of customers with lower spend and income levels.

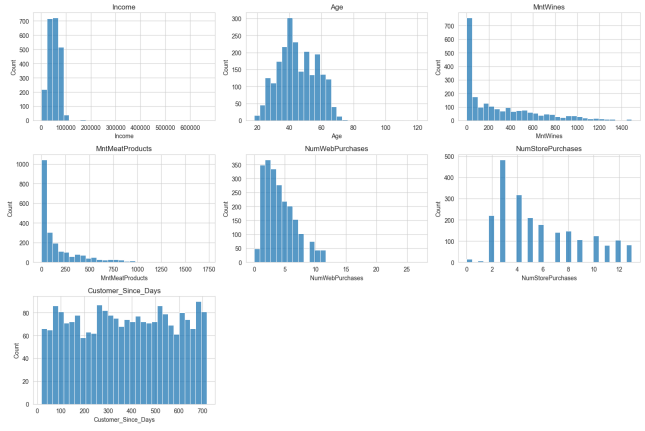

Figure 1: **Key Variable Distributions**

Correlations (see Figure 2) between spending in certain categories and campaign response suggest targeted marketing may be effective.

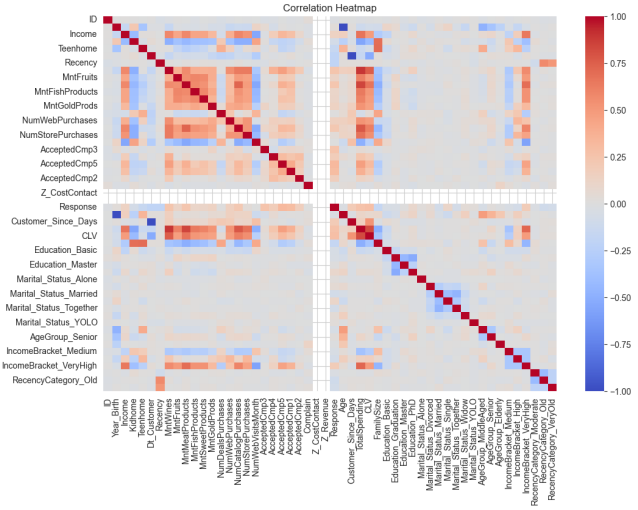

Figure 2: **Correlation Heatmap**

2.3 Data Quality Aspects

There were 24 missing values in the income attribute, which were just removed. Some outliers in spending could represent high-value customers or data errors. Correlation analysis has helped understand relationships without biasing the dataset.

2.4 Visual Exploration of Data Properties and Hypotheses

Histograms and box plots show distributions and relationships between spending behaviors and campaign responses. The Amount Spent on Wines, Meat Products, and other categories is highly right-skewed. Most customers spend low to moderate amounts, with a few high spenders (see **Figure 1**). Campaigns could be designed to target these different segments uniquely. The Number of Web and Store Purchases exhibits a varied distribution, indicating diverse shopping habits that can be leveraged to personalize the marketing strategy.

The count plot (see **Figure 3**) reveals that the sixth campaign had a notably higher acceptance rate than the others. This could imply either a more compelling offer or perhaps a broader targeting strategy. The low acceptance of certain other campaigns suggests a potential mismatch between the offers and the customer segments they were aimed at.

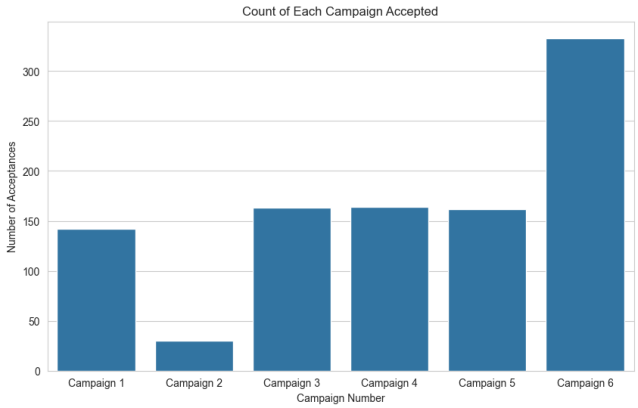

Figure 3: **Count Plot for Accepted Campaigns**

Figure 4 shows key variables plotted against an accepted campaign, whereas **Figure 5** shows key variables against the total number of accepted campaigns. When looking at the Amount Spent on Sweet Products, we can see a more uniform distribution across the campaigns, suggesting that customer preference for sweet products may not be a strong differentiator in campaign acceptance. Customers accepting Campaign 3 seem to have a low amount of purchases on Meat and Wine and also have lower Incomes.

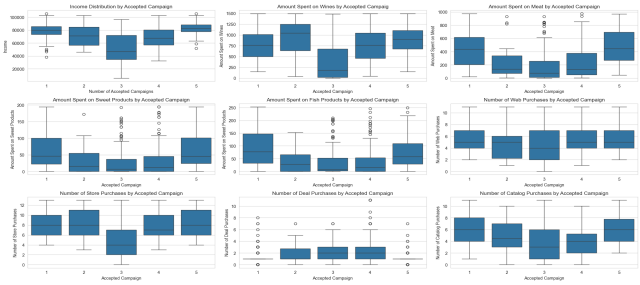

Figure 4: **Key Variables by Accepted Campaign**

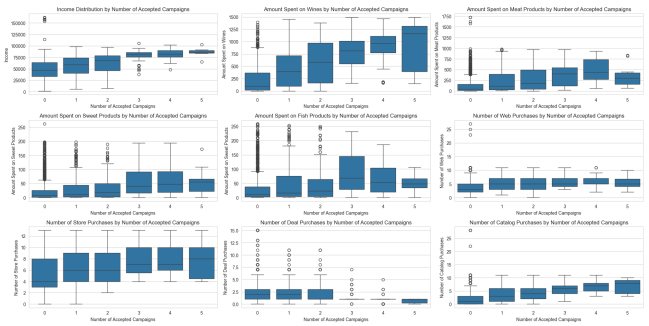

Figure 5: **Key Variables by Total Number of Accepted Campaigns**

Figure 6 shows no major differences between recency categories and number of accepted campaigns.

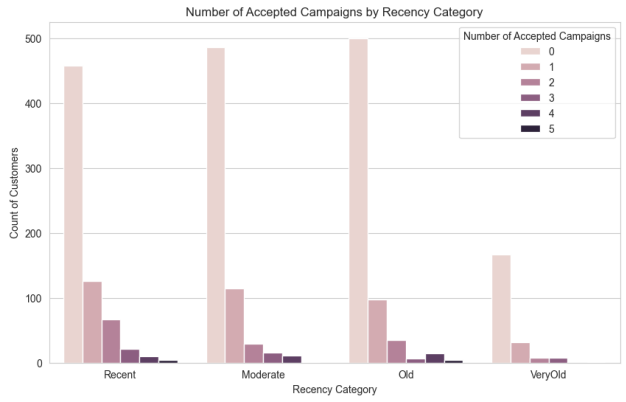

Figure 6: **Number of Accepted Campaigns by Recency (Last Purchase)**

2.5 Ethical Sensitivity and Data Bias

The dataset includes attributes like marital status and education, which could be ethically sensitive. It's essential to avoid discriminatory practices based on these attributes. The data set appears unbalanced with respect to the 'Campaign 6 (Response)' attribute, which could introduce bias in model predictions.

2.6 Potential Risks and Bias

Potential risks include model overfitting and bias in campaign acceptance predictions. Questions for external experts would revolve around the ethical use of demographic data and the representativeness of the dataset.

2.7 Data Preparation Actions

Based on the analysis, actions such as normalization of skewed attributes, encoding categorical variables, and missing values are likely required. Variable Transformation is something that will be viewed closer during the modeling task as different models are sensitive to different transformations and some others are more robust and don't require transformations. Outliers will for now be left untouched.

3 Data Preparation Report

3.1 Derived Attributes Analysis

Analysis showed potential for features like 'TotalSpending' and 'Customer Lifetime Value (CLV)'. These derived attributes could enhance the model's predictive power. `TotalSpending` is the cumulative sum amount of all purchases (Mnt variables). `CLV` is the product of `TotalSpending` and the number of days since a customer has been registered. The number of days was derived by taking the most recent registered customer and adding the `Recency` (number of days since last purchase) and subtracting the `Dt_Customer` variable. `FamilySize` is another derived attribute which is the sum of kids and teens in the household. `IncomeBracket` is the categorizing of Income into brackets corresponding to quantiles. The customers were also divided into Age Groups and Recency Categories.

3.2 External Data Sources Analysis

Hypothetically, integrating external data sources such as economic indicators or regional purchasing power could provide additional context for the model.

3.3 Pre-processing Steps

Pre-processing included:

1. Dropping missing values in 'Income'.
2. Encoding the binary encoded labels (AcceptedCmp1 through 5 and Response which corresponds to Campaign 6) into a single variable AcceptedCampaigns which contains the list of accepted campaigns for each customer
3. Creating derived attributes like 'TotalSpending', 'AgeGroup', `CLV`, `Customer_Since_Days`, `FamilySize`, `IncomeBrakcet`, `RecencyCategory`.