Marketing Campaign Analysis

Viha Mashruwala
Yash Karkhanis
Yashaswini Kulkarni

Team Members:
Aditi Pala
Viha Mashruwala
Yash Karkhanis
Yashaswini Kulkarni

Business Use Case

A retail chain wants to offer Gold Membership to its customers. Gold Membership offers 20% discount on products for an annual fee. They are planning to launch a marketing campaign to advertise the same and would like to use data of the previous marketing campaign to enhance efficiency of this marketing campaign

Objective

The main objective of the project is to create a predictive model to identify the customers, that are likely to respond positively to the marketing campaigns, based on the past campaign data, which will help in enhancing the campaign efficiency.

Identify which customer demographic and behavioral factors influence the likelihood of a customer's positive response. This insight will help in tailoring the marketing strategies and campaigns toward the customer segments most likely to convert to subscription services.

Understanding Our Customers: A Look at the Data

This dataset contains 2240 rows and 22 columns of information about each customer.

Demographics:

- Year_Birth: Age of the customer
- Education: Customer's level of education
- Marital: Customer's marital status
- Kidhome : Number of small children in customer's household
- Teenhome Number of teenagers in customer's household

Family Dynamics:

- ID: Unique identifier for each customer
- Dt_Customer: Date of customer's enrollment with the company
- Complain: Did the customer complain in the last 2 years?
- Response: Did the customer respond positively to the last campaign? (1 for yes, 0 for no)

Income:

- Income : Customer's yearly household income

Website Engagement:

- NumWebVisitsMonth: Number of visits to the company's website in the last month

Recency:

- Recency: Number of days since the last purchase

Spending Habits:

- MntFishProducts: Amount spent on fish products in the past two years
- MntMeatProducts: Amount spent on meat products in the past two years
- MntFruits: Amount spent on fruits products in the past two years
- MntSweetProducts: Amount spent on sweet products in the past two years
- MntWines: Amount spent on wine products in the past two years
- MntGoldProds: Amount spent on gold products in the past two years

Purchase Behaviour:

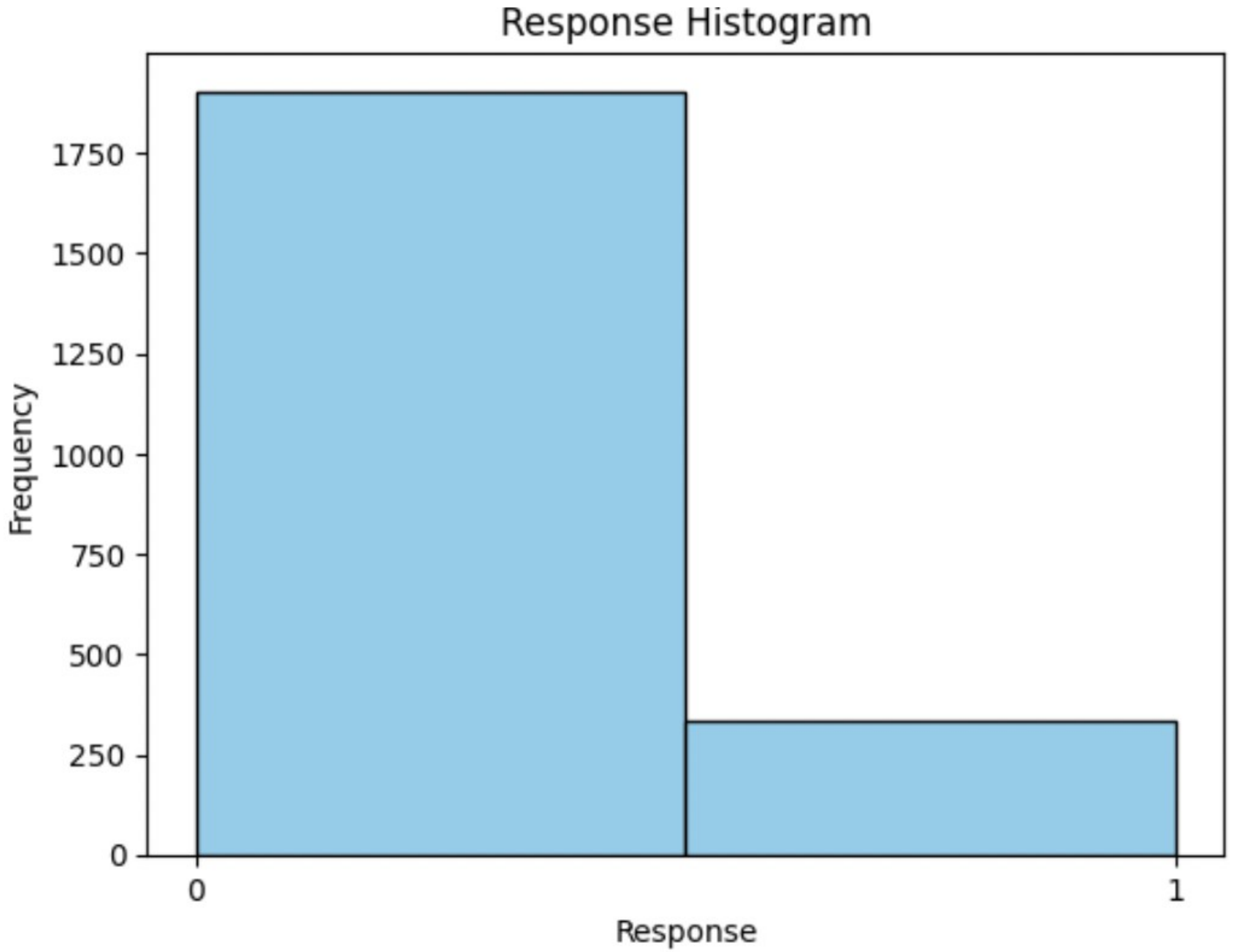
- NumDealsPurchases: Number of purchases made with a discount
- NumCatalogPurchases: Number of purchases made using a catalog
- NumStorePurchases: Number of purchases made directly in stores
- NumWebPurchases: Number of purchases made through the company's website

Data set sourced from Kaggle :
https://www.kaggle.com/datasets/ahsan81/superstore-marketing-campaign-dataset/data

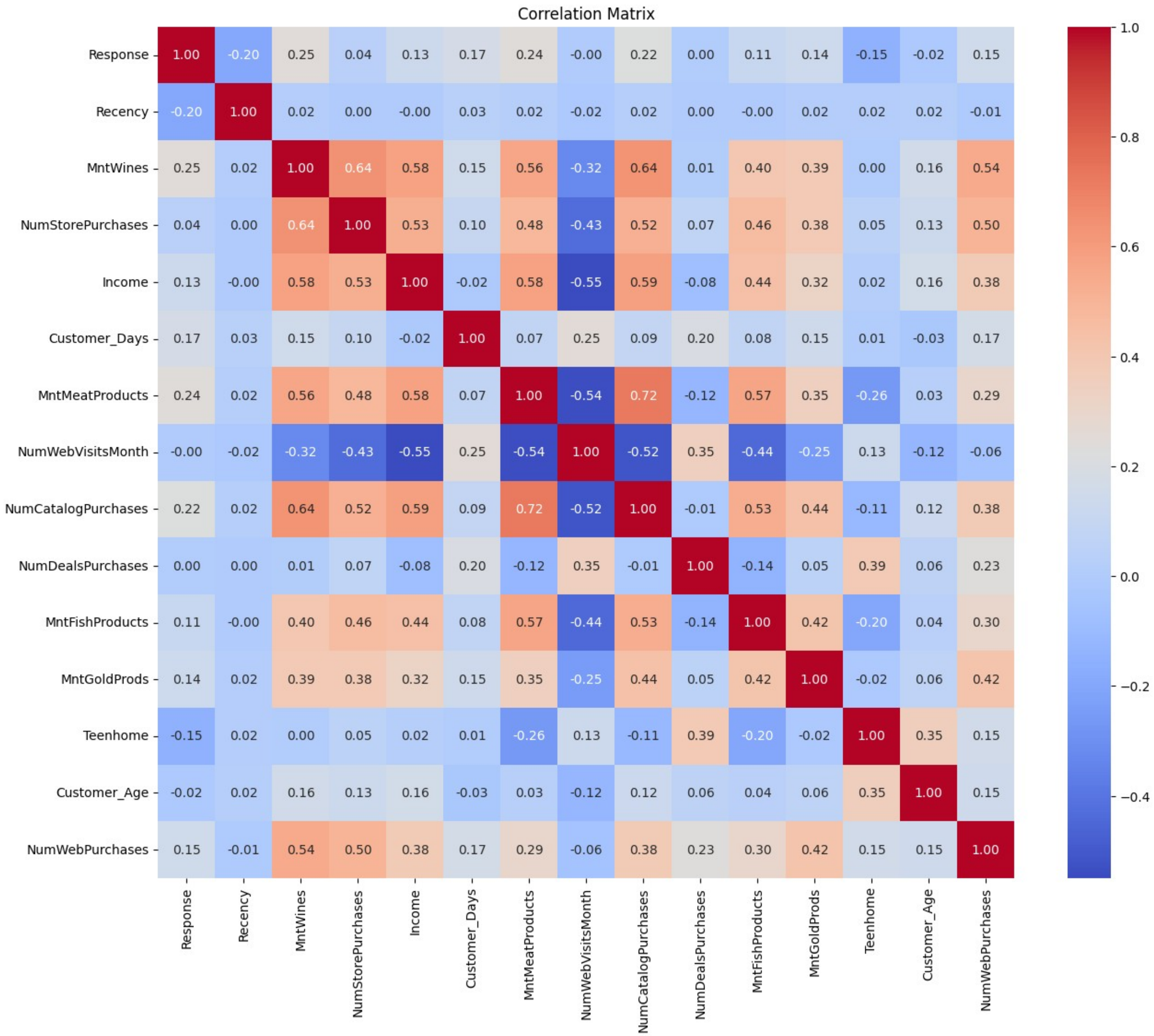
Data Preprocessing: Preparing Data for Analysis

- ***Date Format Conversion:*** Ensured all dates present are in standardized format
- ***Handling Missing Values:***
  - Found 24 missing values in the "Income" column
  - Imputed missing “Income” values using the median for each education level
- ***Categorical Variables Exploration:***
  - Explored "Marital Status" and "Education Level" categories.
- ***Cleaning Marital Status Categories:***
  - Dropped irrelevant categories: "YOLO" and "Absurd" (4 rows).
  - Merged "Alone" with "Single."
- ***Calculating Age and Membership Duration:***
  - Derived customer age and years of membership by subtracting birth and enrollment dates from the assumed data collection year (2015).

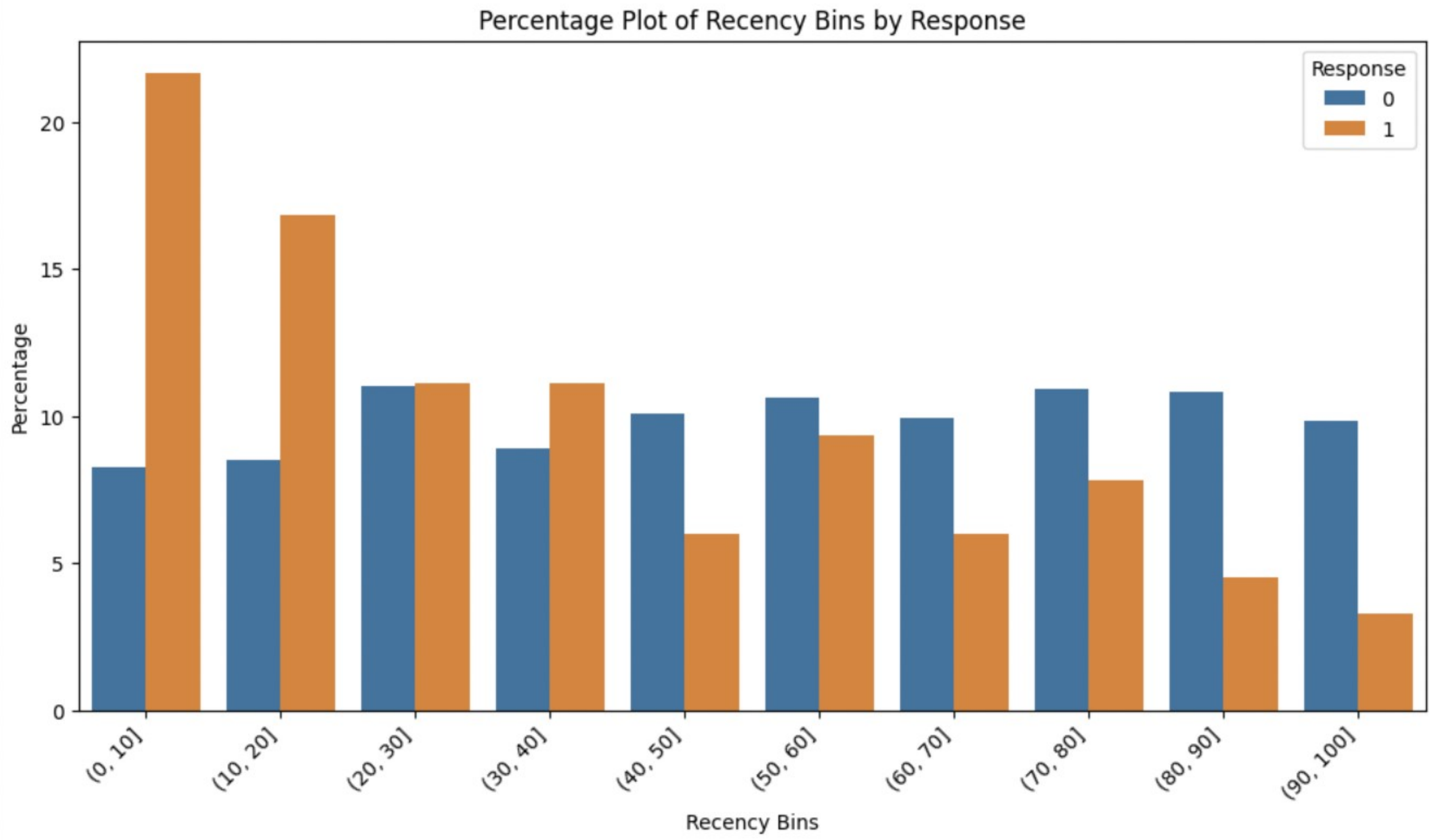
EDA - Distribution of response variable



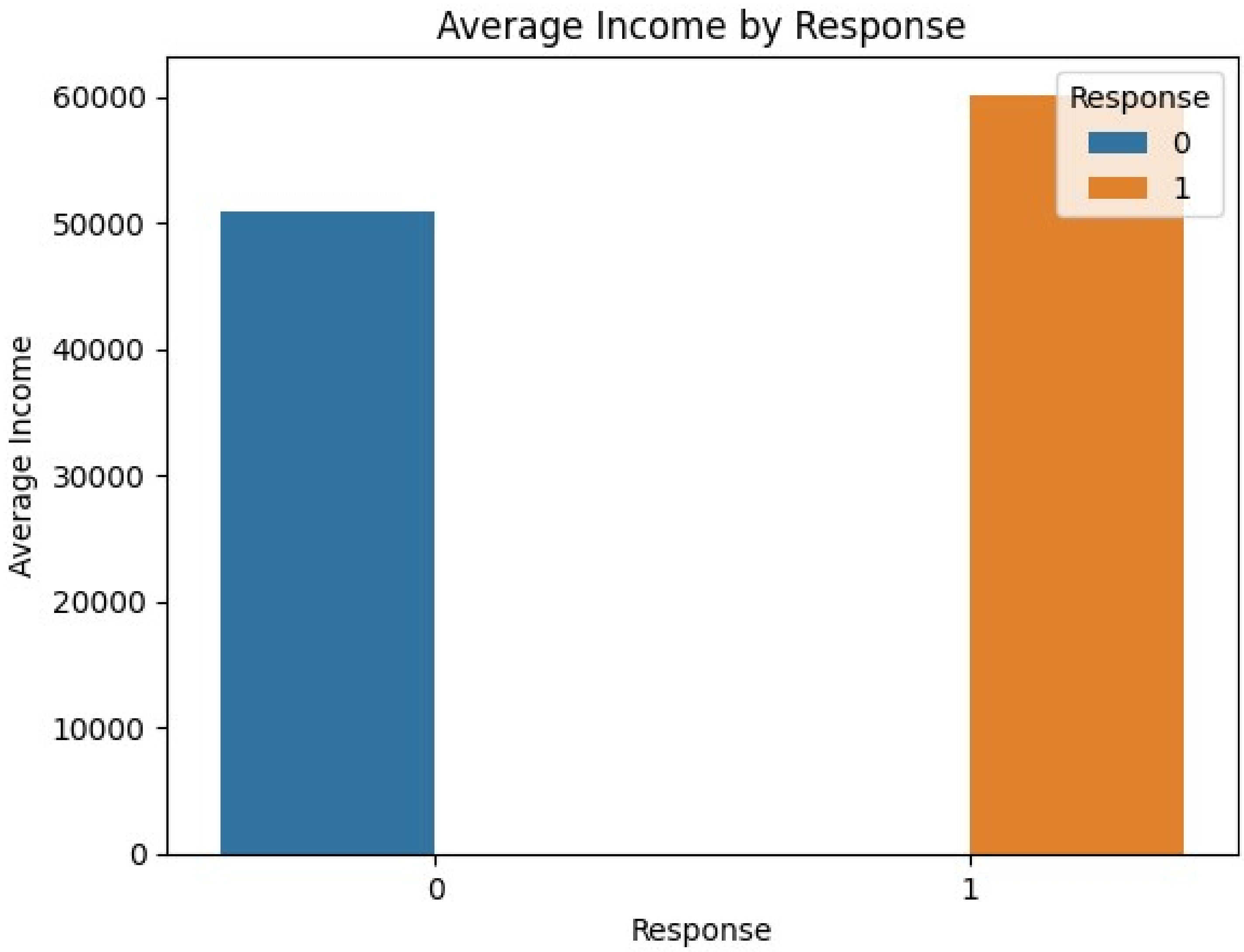
EDA - Correlation Matrix



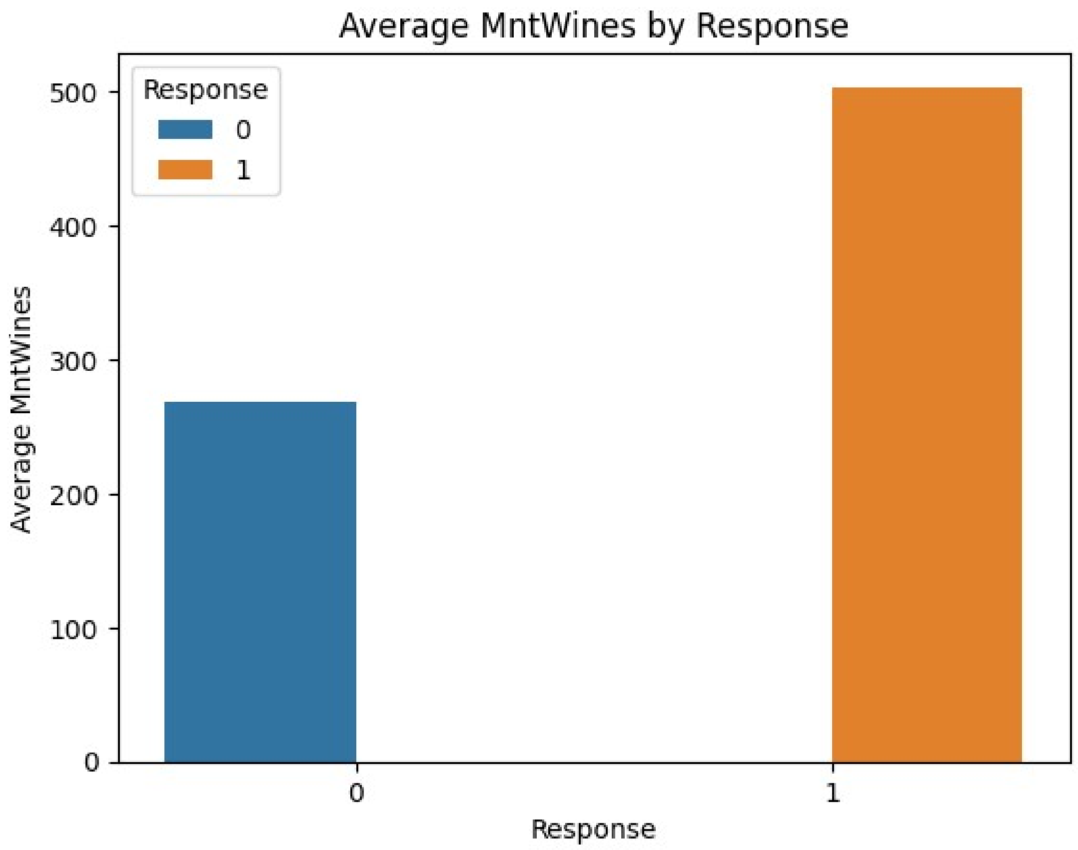
EDA – Percentage plot of Recency by Response



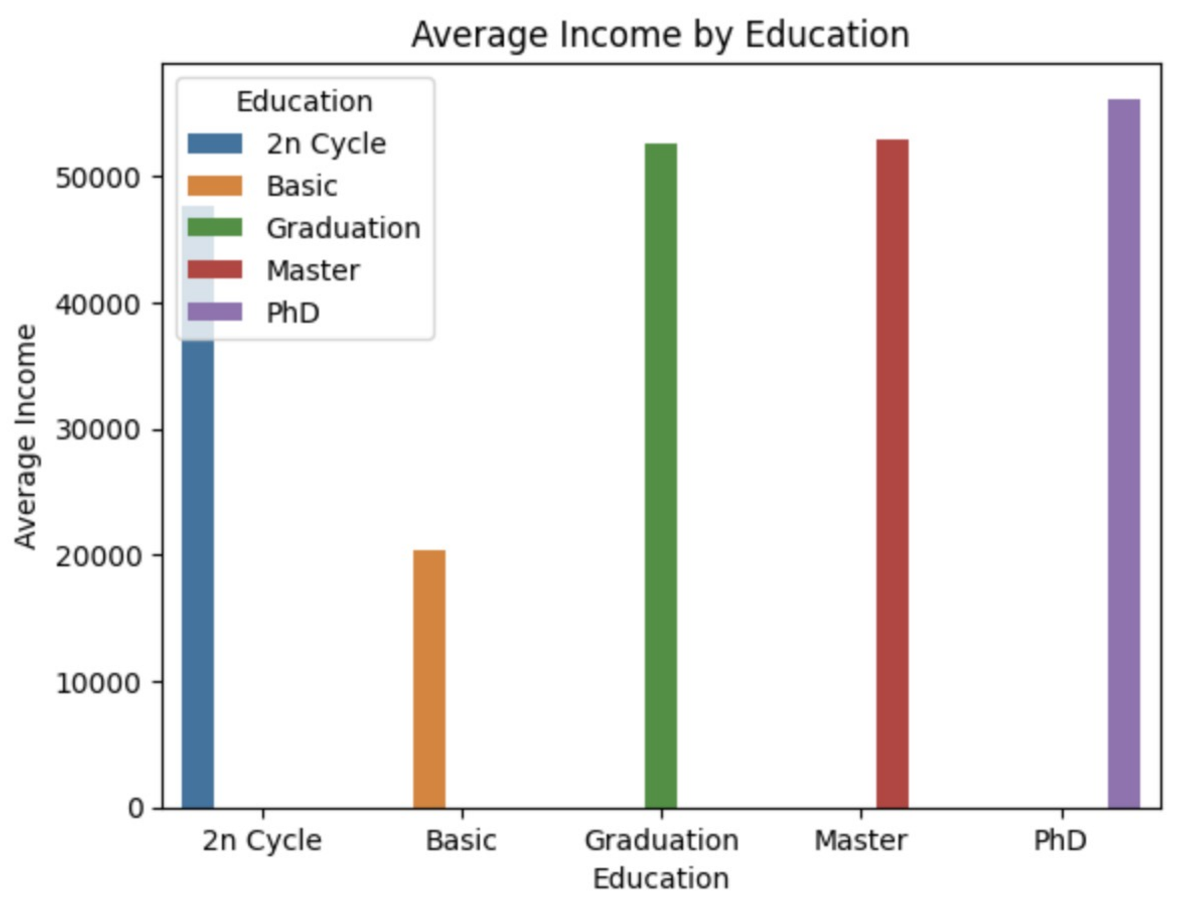
EDA – Average Income by Response



EDA – Average Amount spent on wines by Response



EDA – Average Income based on Education level



Feature Engineering

- One hot encoding on categorical variables [Education, marital status]
- Train test Split :
  - 80-20 train test split
  - Stratified sampling to maintain target variable proportions in train & test data
- Scaling: Standardize the value

Models

- Use same train test split across all models to ensure we get comparative results
- Using AUC as a metric for evaluating model performance
- Recall of Response=1 must be high:
  - Since we are targeting customers who we think will buy the gold membership, we would rather target a few customers who are not likely to buy the gold membership than not reach out to people who are interested in it.

Regular Classification

1. **Logistic Regression**
   - No regularization
   - AUC score : 63.32

2. GBT

- three-fold cross validation with AUC as metric to tune two hyper parameters – max depth and step size.
- Best AUC of 65.89 with max depth = 5, and step size = 0.3.

Weighted Classification – Class Weights

- Models are performing poorly on the imbalanced dataset. Hence we used weighted classification to help improve accuracy

- Since the class distribution is 85% response = 0, and 15% response = 1, we ideally assign class weights to balance them, hence 0.85 to response = 1 and, 0.15 to response = 0

- As mentioned before, the recall of class response =1 is important for our use case, hence we slightly increase the importance of response =1 by assigning the below weights – response 1=0.9 and response 0= 0.1

Weighted Classification-Models

1. Weighted Logistic Regression
   - Same model with class weights = {0 : 0.1, 1 : 0.9}
   - AUC score : 75.15
   - Recall of response = 1 is 83%

2. Weighted GBT
   - Same procedure as before with the class weights = {0 : 0.1, 1 : 0.9}
   - Best AUC of 77.06 with max depth = 5, and step size = 0.3.
   - Recall of response = 1 is 83%

Model Results

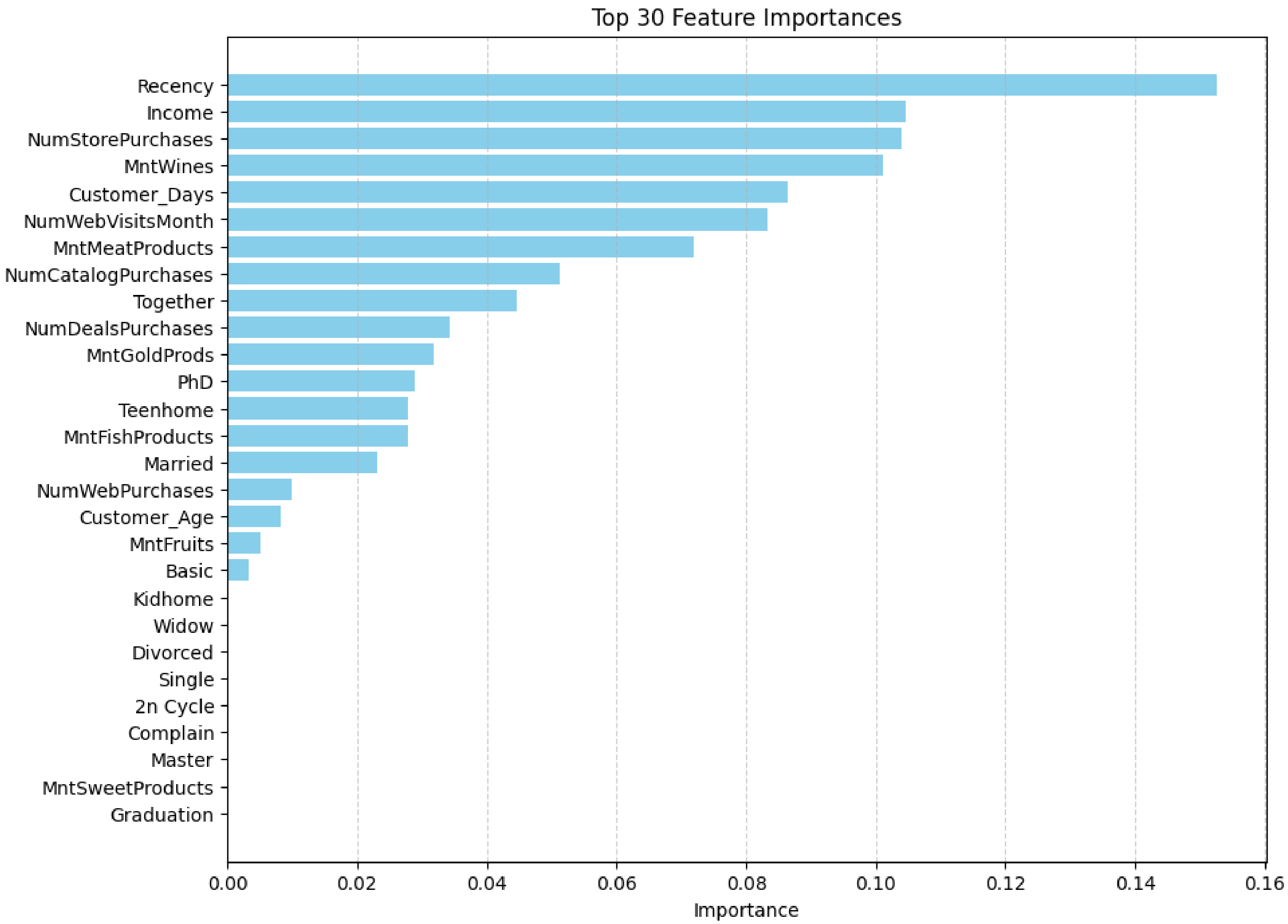
Model	AUC	Recall of response =1
Logistic Regression	63.32	30%
Gradient Boost Tree	65.89	38%
Weighted Logistic Regression	75.15	83%
Weighted Gradient Boost Tree	77.06	83%

Our use case is to rank customers based on their probability of responding positively to the gold membership campaign. Hence an AUC score of 77 and recall of response =1 value of 83% is sufficient

Sample Prediction

Customer_ID	Accept_Prob
1162	0.971500326
6215	0.948906081
5329	0.93108072
4096	0.917169054
7505	0.897373815
6983	0.877917117
2457	0.861664188
7224	0.837589703
2428	0.82888789
11112	0.818803146
3436	0.795263808
4286	0.792598967
10643	0.753116142
1802	0.740926854
2114	0.739952295
968	0.702537685
3153	0.693555913
6504	0.68901981
2495	0.685534838
4261	0.684386919

Feature Importance



Key Insights

Customers who spend more on alcohol and meat products are more likely to buy the gold membership

Customers who visit the store are more likely to purchase gold membership than those who use other buying options