Marketing Campaign Analysis

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 prioritize 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.

Dataset Overview:

- This dataset is marketing campaign data, where each entry is a unique customer and includes various attributes about the customer and whether they responded positively to the campaign. The dataset contains 2240 rows and 22 columns. The attributes in the dataset are:
    1. Response - 1 if customer accepted the offer in the last campaign, 0 otherwise
    2. ID - Unique ID of each customer
    3. Year_Birth - Age of the customer
    4. Complain - 1 if the customer complained in the last 2 years
    5. Dt_Customer - date of customer's enrollment with the company
    6. Education - customer's level of education
    7. Marital - customer's marital status
    8. Kidhome - number of small children in customer's household
    9. Teenhome - number of teenagers in customer's household
    10. Income - customer's yearly household income
    11. MntFishProducts - the amount spent on fish products in the last 2 years
    12. MntMeatProducts - the amount spent on meat products in the last 2 years
    13. MntFruits - the amount spent on fruits products in the last 2 years
    14. MntSweetProducts - amount spent on sweet products in the last 2 years
    15. MntWines - the amount spent on wine products in the last 2 years
    16. MntGoldProds - the amount spent on gold products in the last 2 years
    17. NumDealsPurchases - number of purchases made with discount
    18. NumCatalogPurchases - number of purchases made using catalog (buying goods to

19. be shipped through the mail)
20. NumStorePurchases - number of purchases made directly in stores
21. NumWebPurchases - number of purchases made through the company's website
22. NumWebVisitsMonth - number of visits to company's website in the last month
23. Recency - number of days since the last purchase

Data Preprocessing:

1. Date Format Conversion
   To ensure data consistency, the data formats were standardized. This was implemented to guarantee that all dates within the dataset adhere to a unified and recognizable format.

2. Handling Missing Values
   There were 24 missing values in the "Income" column. The missing values were identified and imputed the median “Income” value corresponding to each education level.

3. Categorical Variables Exploration - Marital Status and Education Level
   In-depth exploration was conducted on the "Marital Status" and "Education Level" categories to identify any inconsistencies in the data. We found several issues with the data in the Martial Status category and the below steps were taken to improve data quality.

   - Dropping Irrelevant Categories: Removed irrelevant categories, namely "YOLO" and "Absurd," accounting for 4 rows of data.
   - Merging Categories: Consolidated the "Alone" category with "Single" as they both represent the same Marital Status.

4. Calculating Age
   Derived age of the customers and how long the person has been a customer of the retail chain because those are better features to work with. This was done by subtracting the date of birth and date since they have been a customer from the assumed data collection year, 2015.

Exploratory Data Analysis:

1. **Distribution of target variable:**
   The image below(fig-1) shows the distribution of response. Here, 0 indicates a negative response, and 1 indicates a positive response.

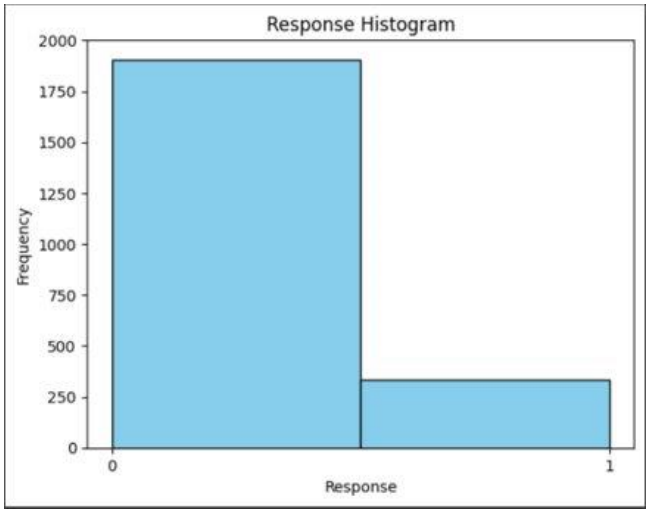


Fig-1

2. **Correlation Matrix:**
   The Correlation Matrix (fig-2) shows the relation between the various variables of the dataset.

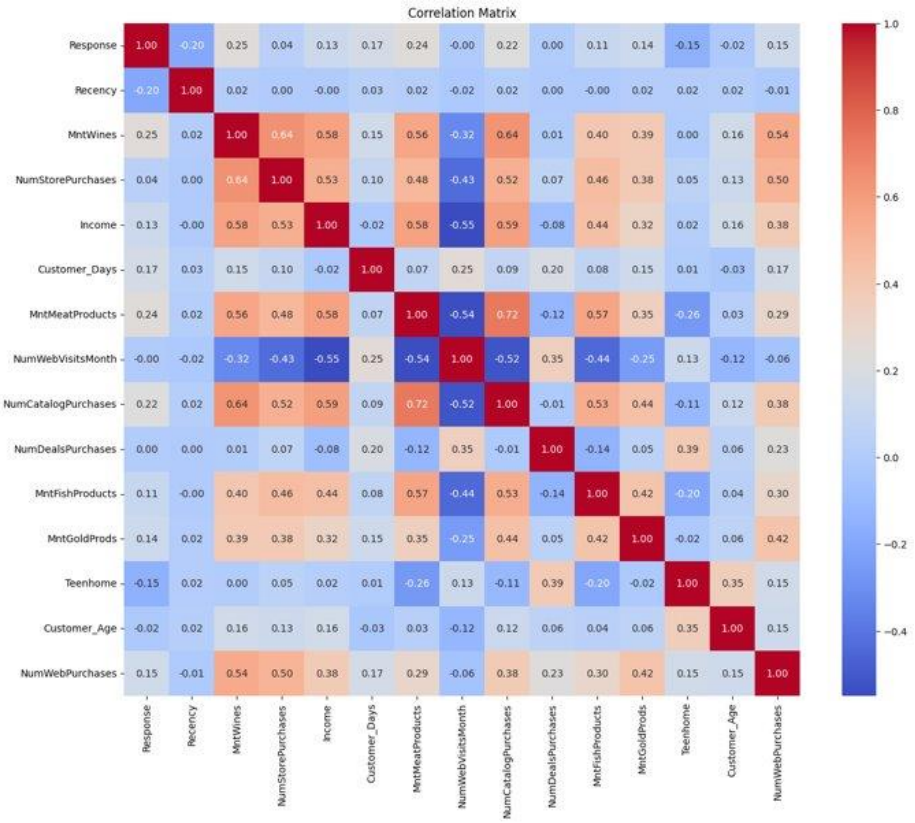


Fig-2

3. Percentage Plot of Recency bins by Response:

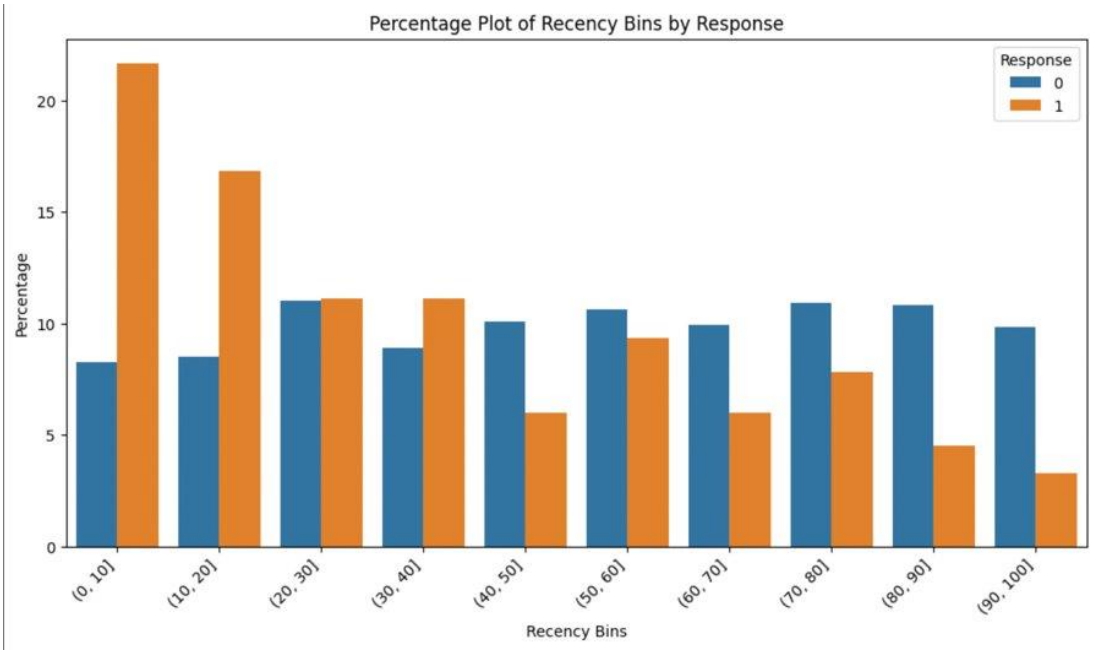


Fig-3

4. Average Income by Response

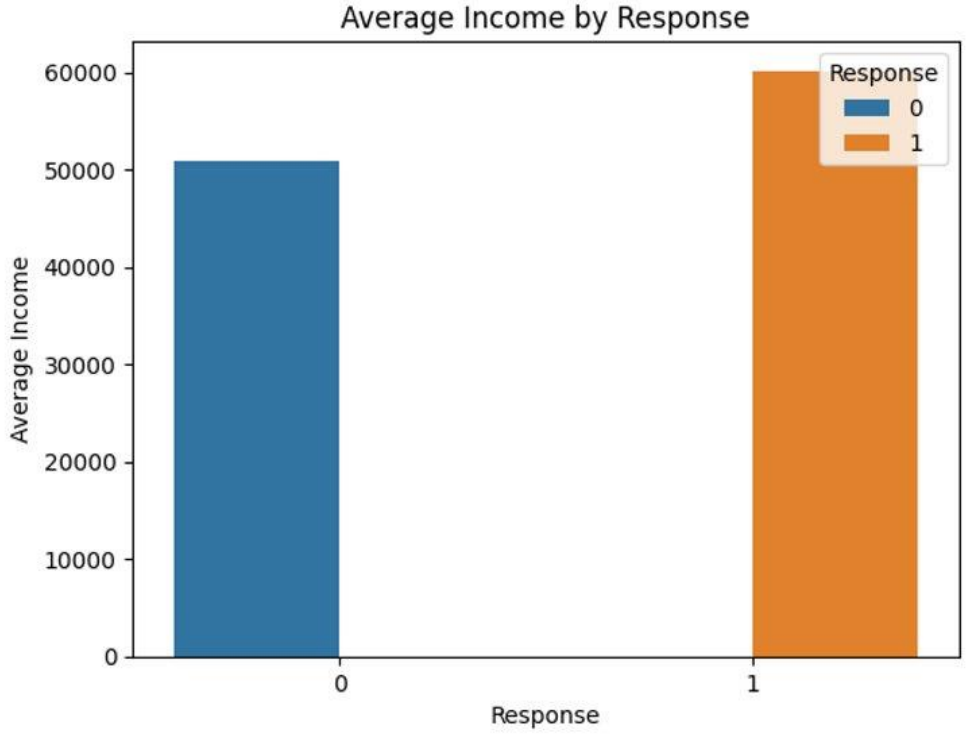


Fig-4

5. Average Amount of Wines by Response

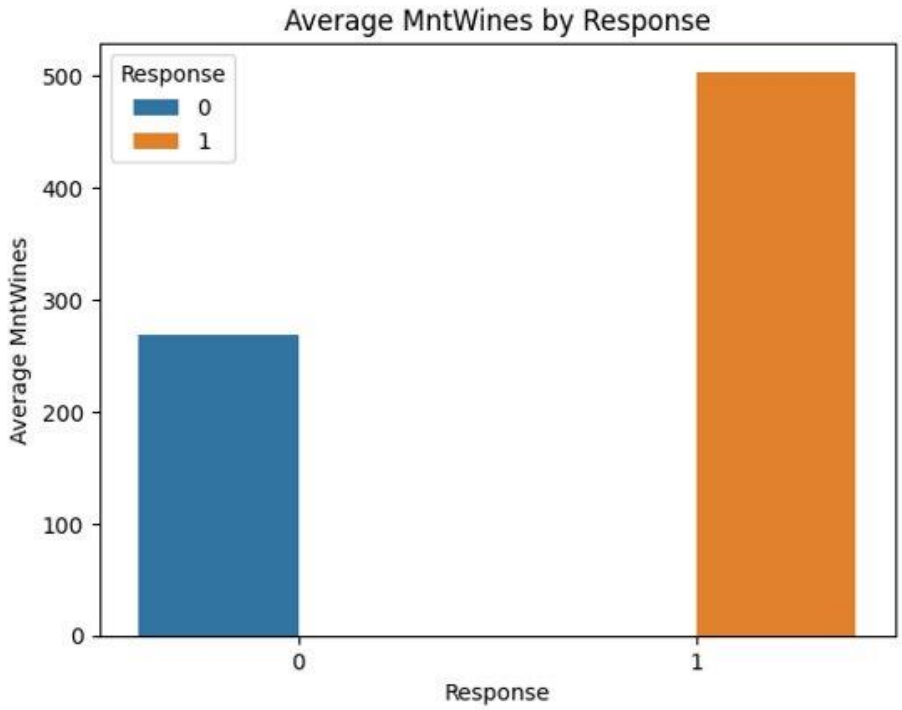


Fig-5

6. Average Income by Education

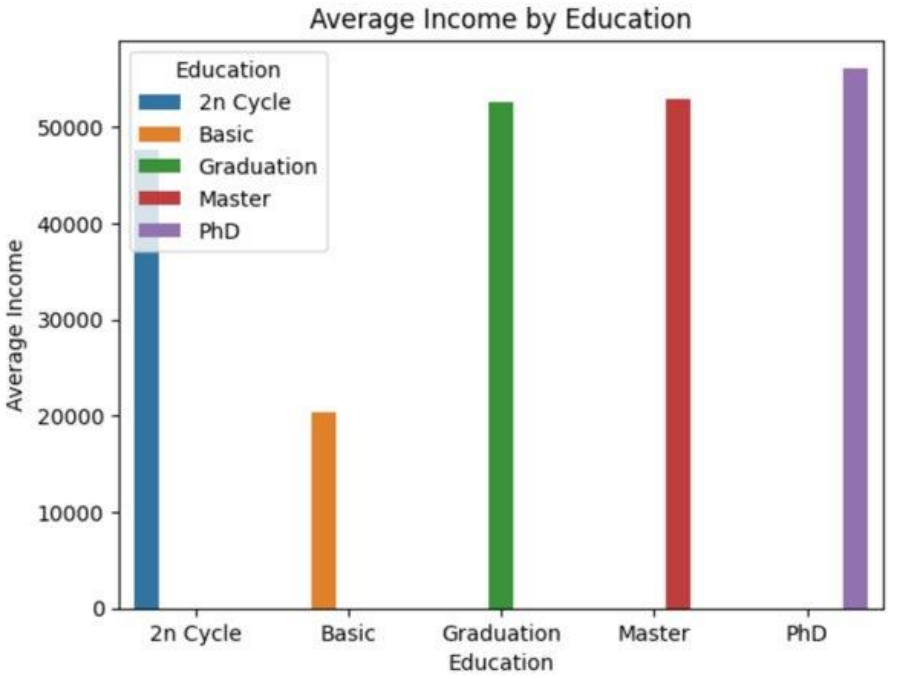


Fig-6

Feature Engineering:

One-Hot Encoding on Categorical Variables

To prepare the categorical variables "Education" and "Marital Status" for inclusion in predictive models, they were one-hot encoded.

Train-Test Split

To evaluate the model's performance effectively the data was split into train and test set. An 80-20 train-test split ratio was used, dedicating 80% of the data for training purposes and reserving the remaining 20% for testing and model evaluation. To ensure the representation of target variable proportions in both the training and testing datasets, a stratified sampling approach was implemented as our target variable is imbalanced.

Scaling

The data was standardized using the standard scaler to bring them to a common scale.

Modeling:

We built classifiers to identify which customers are likely to respond positively to the marketing campaign using regular classification models as well as weighted classification models. The same test data was used across all models to enable us to get comparative results. Since our dataset was imbalanced, we used the ROC curve as the metric for evaluation along with the recall of class response=1 as that is most pertinent to our use case.

Classification

For the regular classification we used two models

1. Logistic Regression – We trained a logistic regression model with no regularization. The model achieved an AUC score of 63.32%. The recall was response=1 was 30%
2. Gradient Boosting Trees – We trained a GBT classifier and used 3-fold cross validation to tune two hyperparameters max depth and step size. The model achieved an AUC of 65.89% with max depth 5 and step size 0.3.

Weighted classification

The above results show the models were not performing well on the imbalanced dataset especially with respect to recall of repones=1. To overcome this, we use weighted classification. We assign class weights to the class to balance the class importance. Our target variable had a distribution of 85% of class 0 and 15% of class 1. To balance this we would ideally assign a class weight of 0.85 to class 1 and wight of 0.15 to class 0. As mentioned above, the recall of class 1 is important for our use case and hence we skew the importance slightly in favor of class 1 by assigning weight of 0.9 to class 1 and 0.1 to class 0. We then rerun both the above models using these class weights.

3. Weighted Logistic Regression – We trained a weighted logistic regression model with the above-mentioned weights and no regularization. The model achieved an AUC score of 75.15%. The recall was response=1 was 83%
4. Gradient Boosting Trees – We trained a GBT classifier with the above-mentioned weights and used 3-fold cross validation to tune two hyperparameters max depth and step size. The model achieved an AUC of 83% with max depth 5 and step size 0.3.

Comparison

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%

From the results we can see that weighted classification models significantly outperformed the regular classification models. Weighted classification resulted in the ROC scores increasing by 10% and Recall increasing by almost 50%. The best performing model was the weighted GBT and hence we use that as the final model.

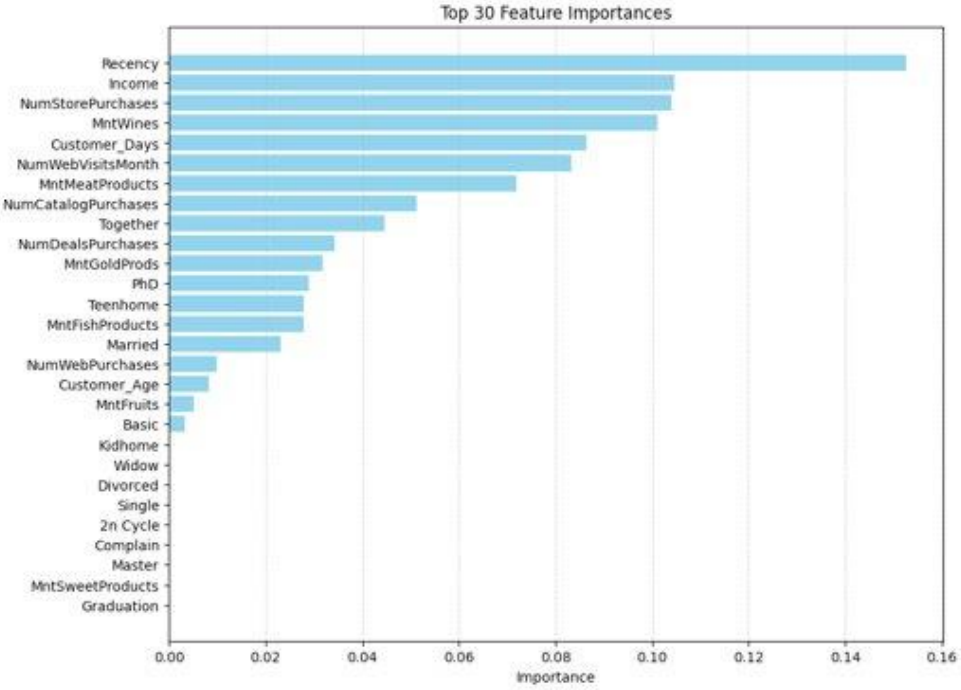
For our use case of optimizing a marketing campaign we believe that an AUC score of 83% is sufficient to achieve the intended result.

Sample Prediction

Given below is an example of the how the model prediction looks like. It ranks the customers from the heights to the lowest probability of responding positively to the marketing campaign.

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



The above graph shows the most important features from the gradient boosting tree model. The top 5 most important features are-

1. Recency
2. Income
3. Number of In-Store Purchases
4. Amount of Wines
5. Customer Days

Key Actionable Insights

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

   Harnessing consumer insights, the retail chain can elevate its advertising strategy by customizing campaigns for those inclined towards alcohol and meat purchases. Introducing exclusive perks, such as a $10 discount on chicken with the gold membership, aligns promotional efforts with identified spending patterns, creating a more enticing proposition for this specific customer segment. This targeted approach not only maximizes the appeal of the gold membership but also cultivates customer loyalty through personalized incentives.

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

    Recognizing that in-store customers display a higher inclination towards purchasing gold memberships offers a strategic advantage in budget allocation. Focusing a significant portion of the marketing budget on impactful in-store advertisements capitalizes on this trend, maximizing the potential for membership conversions. Subsequently, a targeted online portal campaign can further engage potential members, with the remaining budget allocated to catalogue ads ensuring a well-rounded approach that optimally utilizes resources and channels for enhanced membership acquisition.