Machine Learning Group 9 Project: Customer Personality Analysis

Group 9:

Arvind Shankar

Amy Lock

Faye He

Guanyi Lu

Karen Cai

Zahra Yousefpour

Overview

In this analysis, we have identified key customer traits and preferences, segmented our customer groups based on purchasing habits, and pinpointed our top-performing customers. By leveraging this data, we also figured the best predicting model to predict future campaign responses to optimize our business strategies.

Problem Statement

Customer Personality Analysis is a detailed analysis of a company's ideal customers. It helps a business to better understand its customers and makes it easier for them to modify products according to the specific needs, behaviors and concerns of different types of customers.

Customer personality analysis helps a business to modify its product based on its target customers from different types of customer segments. For example, instead of spending money to market a new product to every customer in the company's database, a company can analyze which customer segment is most likely to buy the product and then market the product only on that particular segment.

Our main objective is to gain insights into the personality traits, buying behavior of customers, and campaign response and use this information to make data-driven business decisions. Specifically, we want to answer the following questions:

- What are the most common personality traits among our customers?
- Who are our customer groups and what are the purchasing habits of each group? Who are our best customers?
- Can we use customer personality data, past purchase, and campaign response to predict if a customer is likely to respond to an upcoming campaign?

Dataset

Data is provided by the following link:
https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis

Data Content and Description

People

- 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

Products

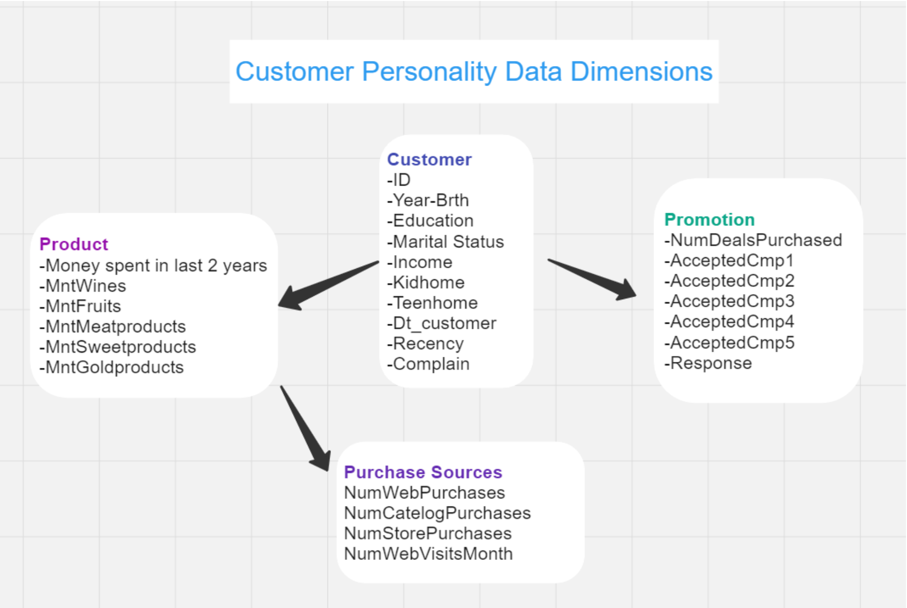
- MntWines: Amount spent on wine in last 2 years
- MntFruits: Amount spent on fruits in last 2 years
- MntMeatProducts: Amount spent on meat in last 2 years
- MntFishProducts: Amount spent on fish in last 2 years
- MntSweetProducts: Amount spent on sweets in last 2 years
- MntGoldProds: Amount spent on gold in last 2 years

Promotion

- NumDealsPurchases: Number of purchases made with a discount
- AcceptedCmp1: if customer accepted the offer in the 1st campaign, 0 otherwise
- AcceptedCmp2: if customer accepted the offer in the 2nd campaign, 0 otherwise
- AcceptedCmp3: if customer accepted the offer in the 3rd campaign, 0 otherwise
- AcceptedCmp4: if customer accepted the offer in the 4th campaign, 0 otherwise
- AcceptedCmp5: if customer accepted the offer in the 5th campaign, 0 otherwise
- Response: if customer accepted the offer in the last campaign, 0 otherwise

Place

- 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



Exploratory Data Analysis - Trends In Our Data

As the first step, we have taken a closer look at separate features in the dataset and performed data cleaning. We dropped “Z_costContact” and “Z_Revenue” columns since they contained the same value over all samples, and therefore there is no impact on our model building. We also dropped the ID column because it has no significant meaning for our clustering approach.

Then, we have determined the main features that we used for analysis by utilizing feature engineering. As a result, there were four main categories of customer personality that we have processed for data visualization, which are *Customer Information, Customer Buying Power, Promotion* and *Customer Purchasing Habit*.

Customer Information

a. ***Age***: We used features ‘Year’ and ‘Year_Birth’ to calculate customer’s ages. *Fig. 1* was plotted to show that the majority of customers are between the age group of 45 to 55. Also, there were some outliers (customers age greater than 120) shown on the graph which have been dropped for further analysis.

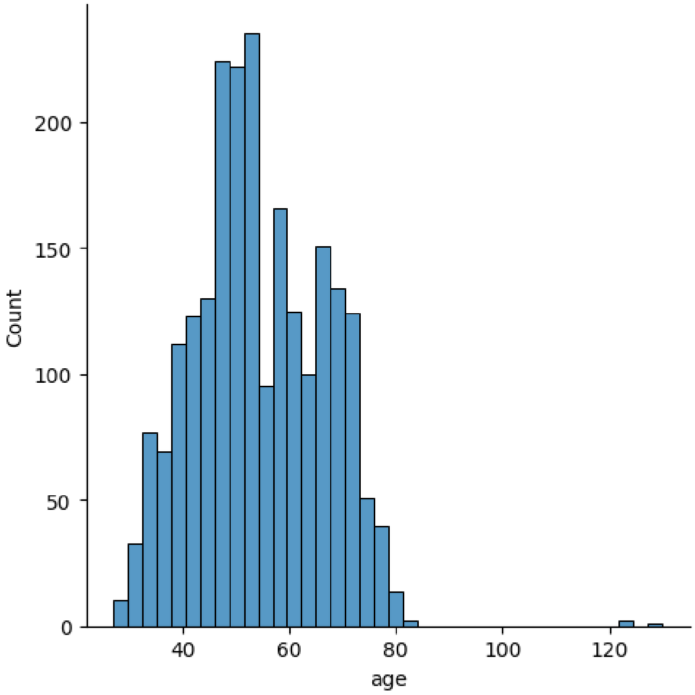


Fig 1. Total Number per Age Group

b. ***Education***: *Fig.2* also has been plotted to show the count for each education group, and it shows that most customer's education level is graduation.

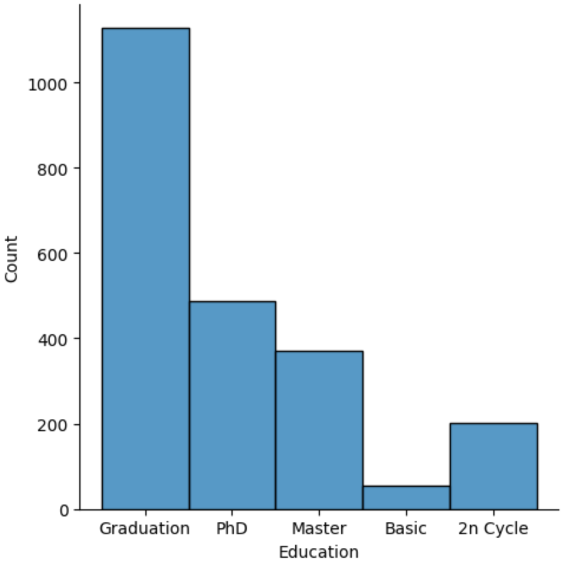


Fig 2. Total Number per Education Level

c. ***Marital Status***: *Fig.3* indicates that the majority of costumes have a relationship of either Married or Together.

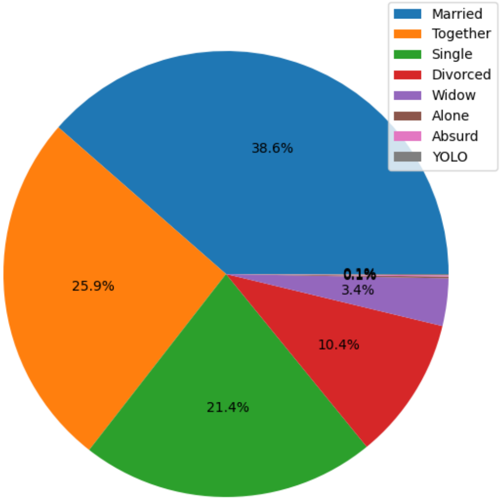


Fig 3. Percentage of Marital Status

d. ***Income***: *Fig.4* shows that most of the customers have salaries below 100k. We can assume that most of the income data follow a normal distribution.

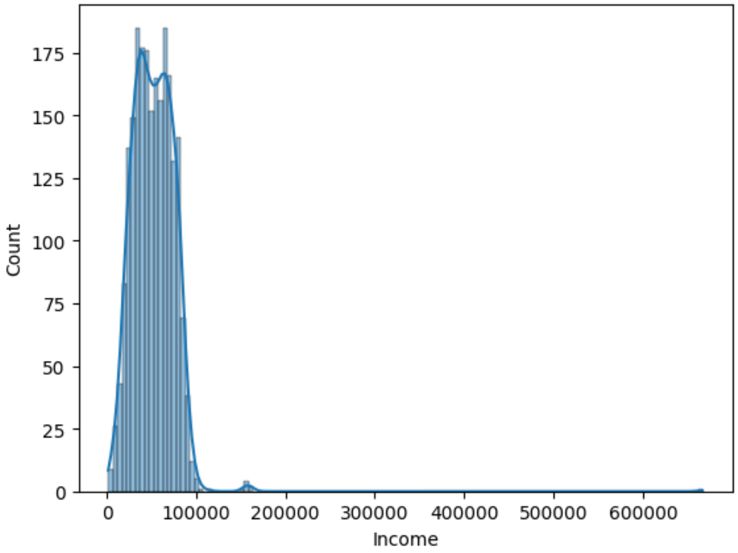


Fig. 4 Total Count of Income Level

e. ***Parental***: We used ‘Kidhome’ and ‘Teenhome’ to plot *Fig.5* and *Fig.6*. Both figures show that most customers don’t have kids or teens home. We have combined these columns to a new column ‘is_parent’ for further analysis.

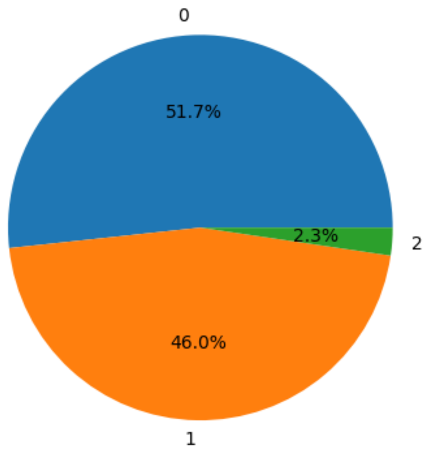


Fig.5 Percentage of Children in Household

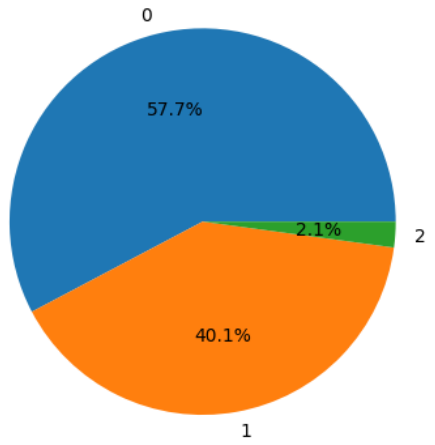


Fig.6 Percentage of Teenages in Household

Customer Buying Power (Total amount of money spent on products)

Fig. 7 shows the total amount of money spent on each product, and it tells us that most customers spend their money on wines and meat products.

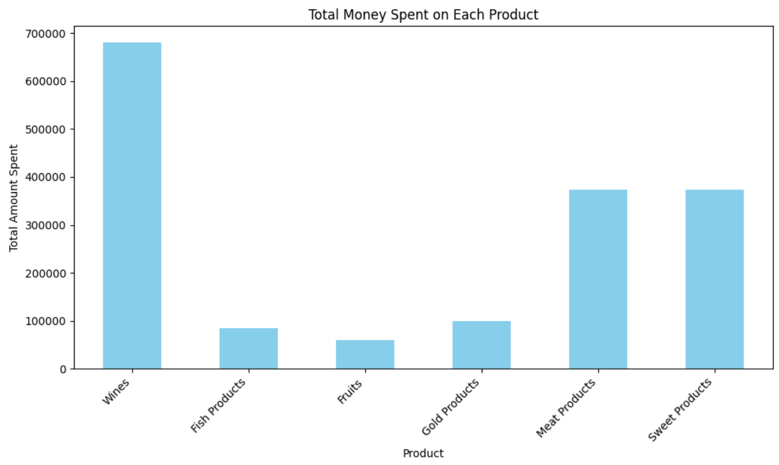


Fig.7 Total Money Spent on Each Product

Customer Purchasing Habit (Online or In-store Purchase)

From *Fig.8* we can conclude that most of the customers were tending to make in-store purchases instead of from online.

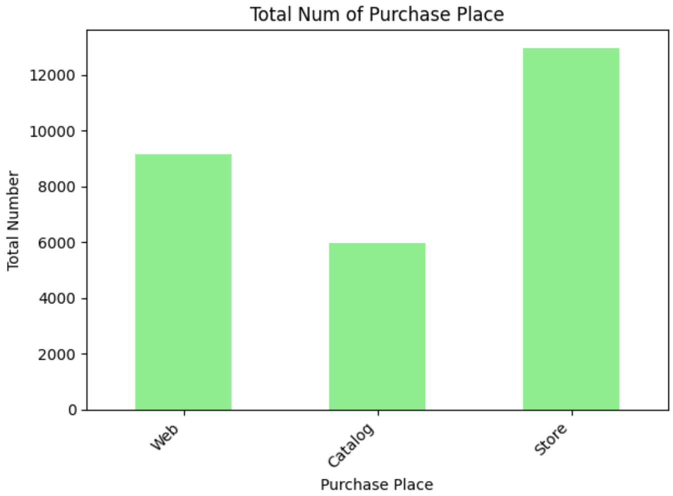


Fig.8 Total Number of Purchase Place

Promotion (Number of offers customer accepted)

The dataset has captured a total of six campaigns, and *Fig.9* shows that the last campaign has the highest number of acceptance from customers.

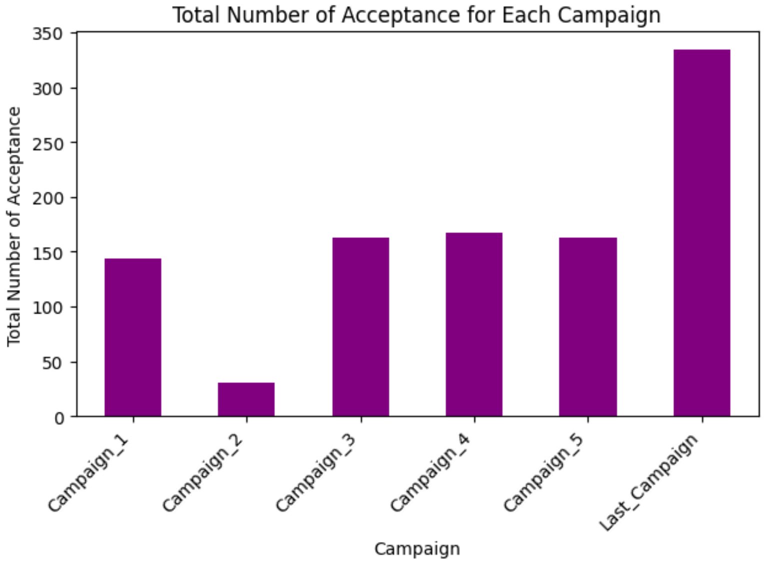


Fig.9 Total Number of Acceptance for Each Campaign

Clustering - Customer Segments

In order to understand our customers better, we used clustering techniques to classify our customers into groups with similar traits and purchasing habits. This information will help us plan and target campaigns that are more likely to speak to each type of customer. It also helps us identify our most valuable customers

Feature Engineering

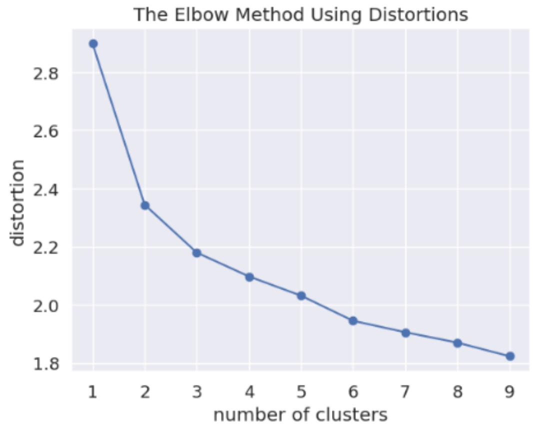
After exploring the data, we decided to use feature engineering to simplify our data set so that we had easier to understand customer segments. We wanted to combine similar data points to improve our predictive abilities and make our results more understandable.

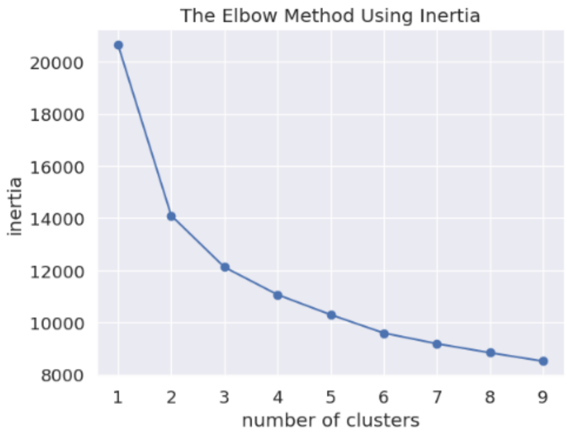
Our changes included:

- Combining ‘Kidhome’ and ‘Teenhome’ into a binary column ‘is_parent’
- Simplifying ‘Marital_status’ which had 8 possible responses to a binary column
- Simplifying ‘Education’ which had 5 possible responses to a binary column
- We combined the 5 columns listing the amount spent on particular items into one column for the total amount spent.

Clustering Process

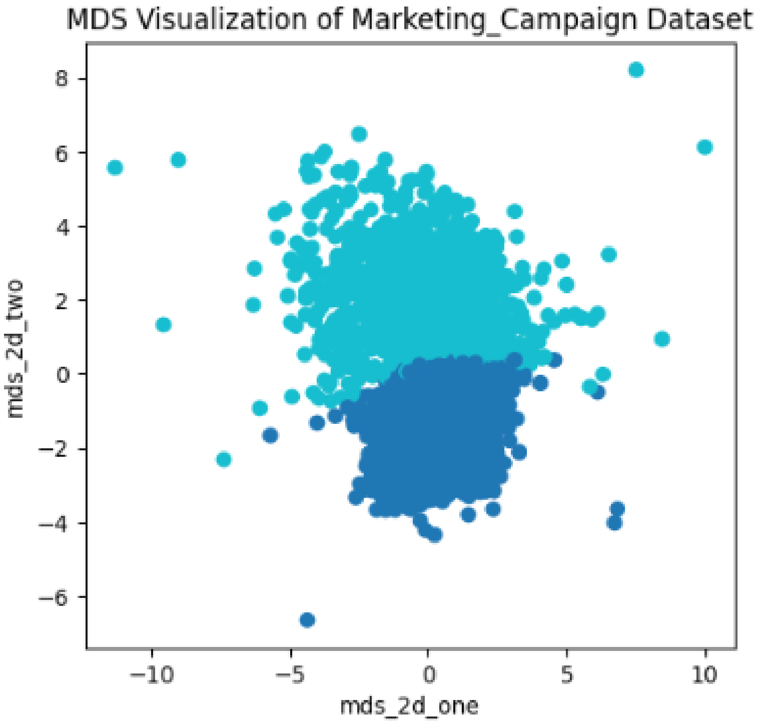
We started our analysis exploring how many groups we should divide our customers into. We used the elbow method with both distortions and inertias to find the best number of clusters and both analysis methods showed two as the optimal number



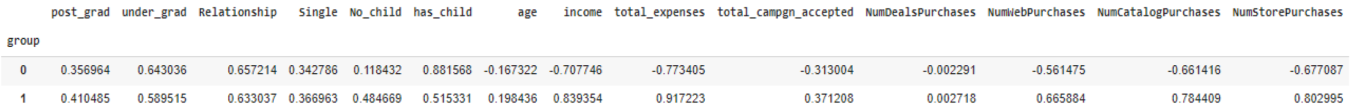


.

We tried a few methods to cluster our customers with similar results from all. Here I will show our results from Multidimensional Scaling but we had very similar groups with K-means and Agglomerative Clustering. We see from the visualization below that the groups are clearly defined and distinct from each other.



The distinctness of the two groups can also be seen when looking at the averages for each column for the two groups (note that these values have been scaled using the Standard Scaler). When looking at the numbers we see that most columns have clear differences between the groups. These two distinct customer groups give us insight into creating and targeting campaigns.



	post_grad	under_grad	Relationship	Single	No_child	has_child	age	income	total_expenses	total_campgn_accepted	NumDealsPurchases	NumWebPurchases	NumCatalogPurchases	NumStorePurchases
group														
0	0.356964	0.643036	0.657214	0.342786	0.118432	0.881568	-0.167322	-0.707746	-0.773405	-0.313004	-0.002291	-0.561475	-0.661416	-0.677087
1	0.410485	0.589515	0.633037	0.366963	0.484669	0.515331	0.198436	0.839354	0.917223	0.371208	0.002718	0.665884	0.784409	0.802995

Clustering Results

Our two customers groups are:

1. **High End frequent Shoppers** : These people are older with high incomes. They buy often from all channels (web, catalog, and in store) but are not particularly driven to buy by deals on products. They spend more and respond to campaigns more than the average customer. These are top customers that should be targeted often.
2. **Young Family, Tight Budget Shoppers**: These customers have children and low incomes. They spend less even when deals are offered, respond to less campaigns, and make less purchases across all three channels. These customers should be targeted thoughtfully with campaigns that appeal to young low income families being thoughtful of the lower expected conversion rate for any paid campaigns.

Supervised Machine Learning Model - Predicting Who Will Convert

In order to target customers thoughtfully for future campaigns, we built a machine learning model that will predict who will likely convert based on customer profiles as well as past purchases and campaign responses. This can be used to send targeted campaigns with higher conversion rates. This will be especially useful with paid campaigns to maximize ROI.

To evaluate our models we looked at the F1 scores as our main performance measure. Since our data was imbalanced with much more non-purchases that purchase the F1- score gives the most meaningful information.

Feature Engineering

We worked with models using three different feature sets in order to compare model performance.

The three feature sets we tested were:

1. all the original columns with minimal changes,
2. engineered an age and customer_since column to get more meaning for the original year_birth and DT_customer columns
3. the same engineered columns used for the clustering

We found the second feature set to give us the best results therefore this is the one we used for our final model.

Modeling Process

After cleaning our data, we looked into if there were any opportunities to reduce our dimensions. Through this analysis, we found that there was little benefit to reducing our dimensions. We do not have a large number of columns so there was little to gain from dimensionality reduction.

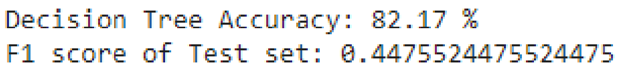
We performed Hyperparameter Tuning on 5 different models and finally determined the winning model based on their performance.

We tested five models (without and without hyperparameter tuning) and here are the results:

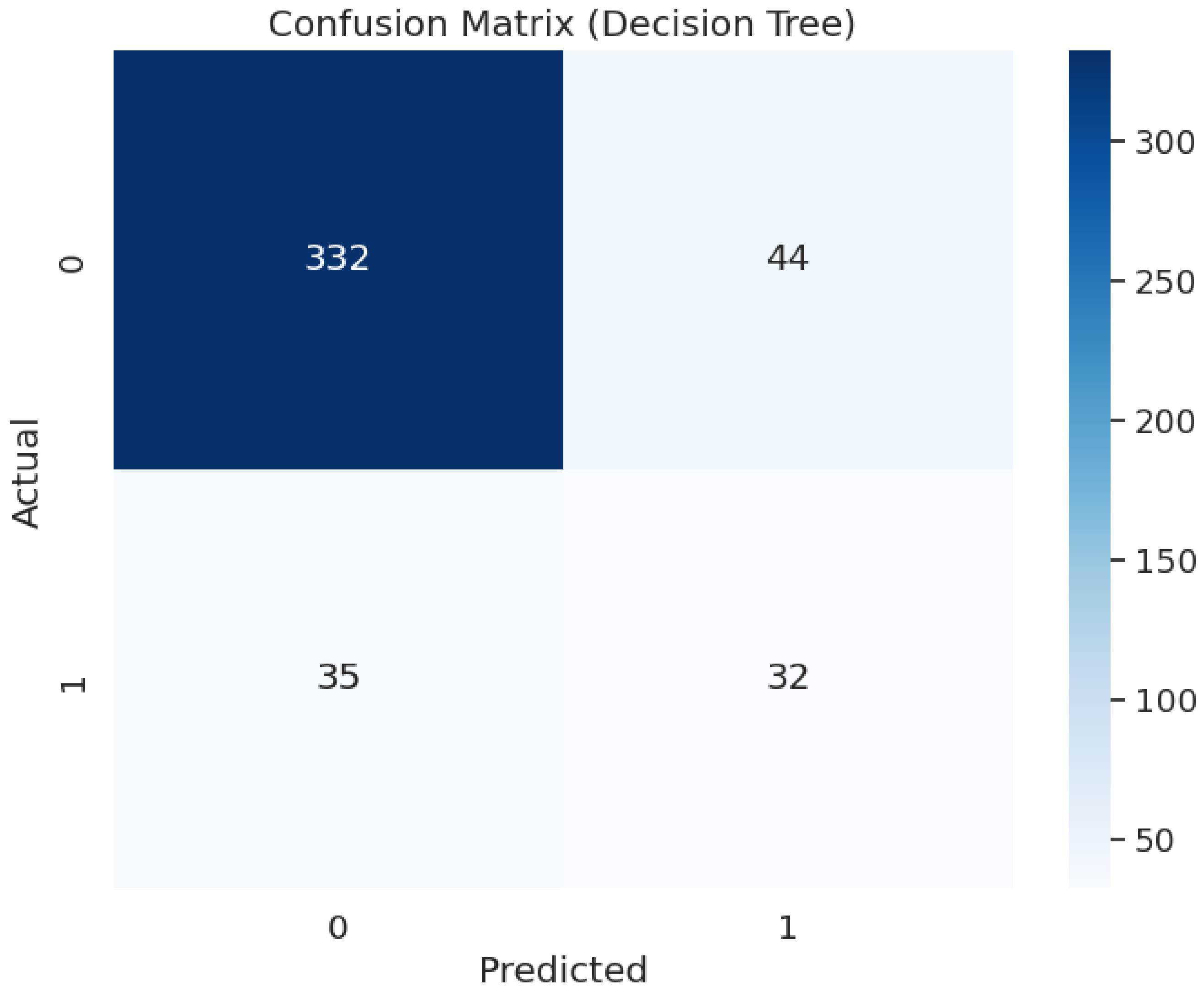
Classification Algorithm	***Without Hyperparameter Tuning***		**With Hyperparameter Tuning**	
	Accuracy Score	**F1 Score**	**Accuracy Score**	**F1 Score**
Decision Tree	82.17%	0.448	86.68%	0.366
Random Forest	88.04%	0.465	88.71%	0.468
Ada Boost	87.58%	0.513	87.36%	0.491
Extra Trees	87.81%	0.471	88.94%	0.515
Gradient Boost	88.04%	0.505	87.13%	0.529

1. ***Decision Tree Classifier*** - This model has the best performance when we run it in the default setting without any tuning; F1 score of 0.448 without tuning vs 0.366 with tuning. Although tuning the model improved the accuracy score, it reduced the F1 score by a huge margin.

   Best Decision Tree Model - *Without Hyperparameter tuning*

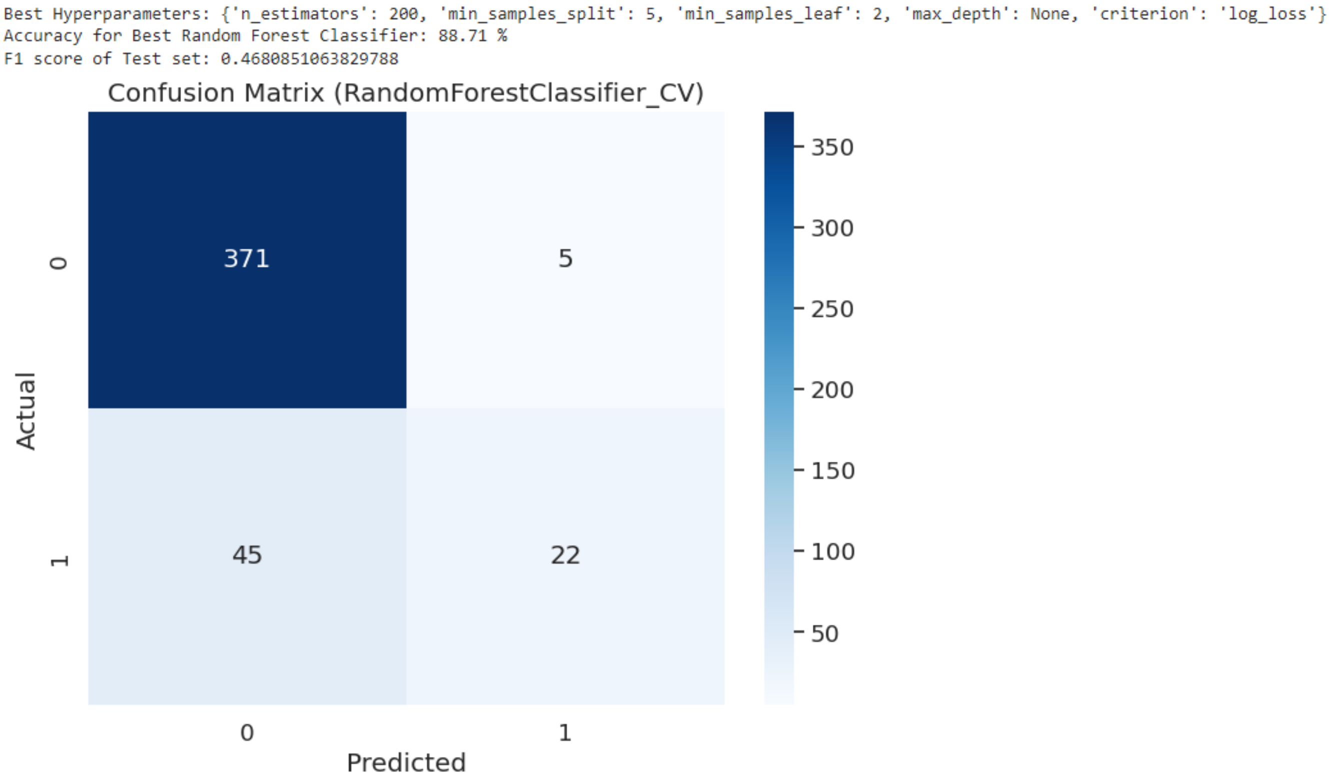


```
Decision Tree Accuracy: 82.17 %
F1 score of Test set: 0.4475524475524475
```



2. ***Random Forest Classifier* -** This algorithm performed comparably similar both with and without tuning, although the tuned model displayed a slightly better performance with improved accuracy and F1 scores

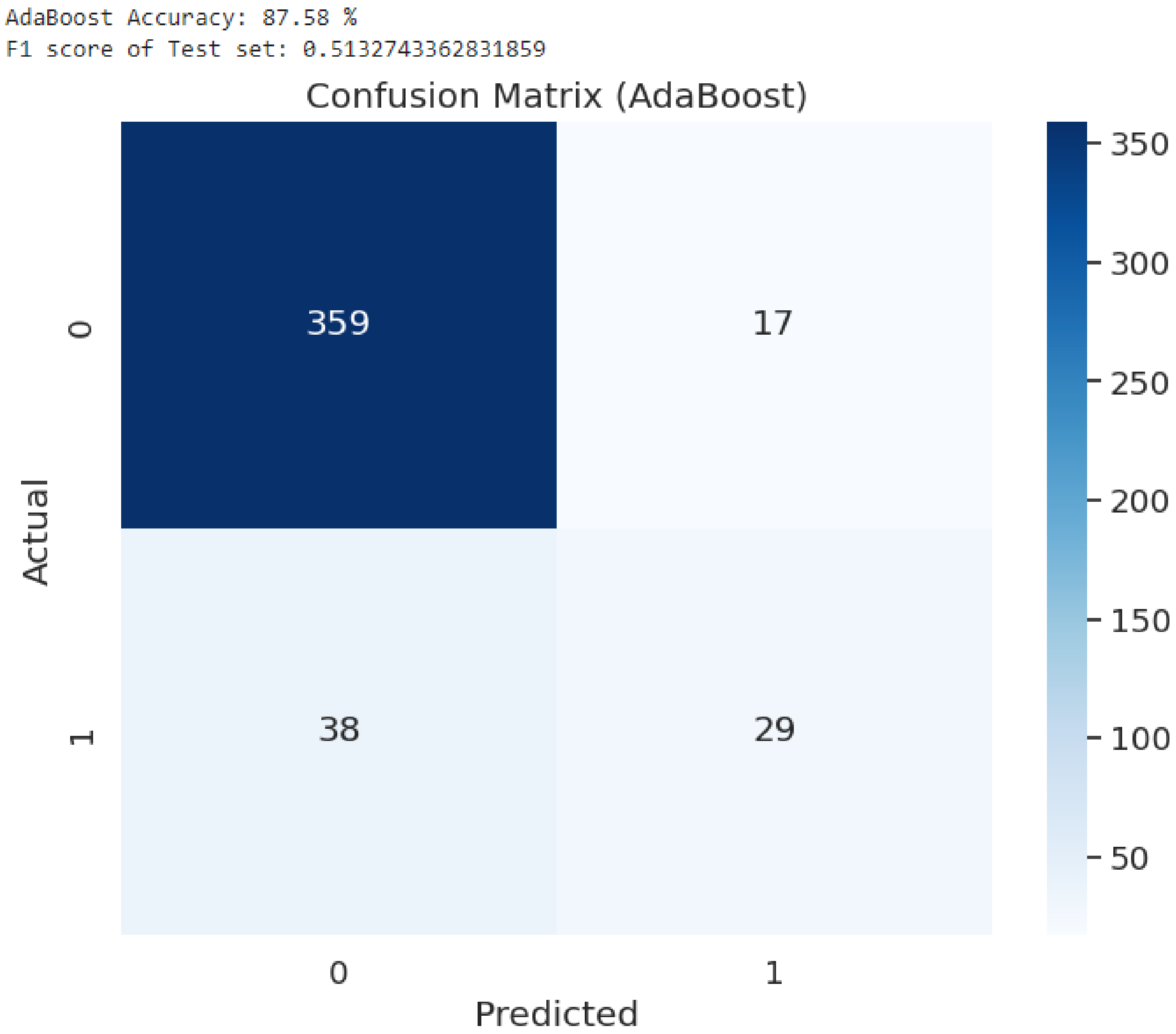
   Best Random Forest Model - *With Hyperparameter tuning*



```
Best Hyperparameters: {'n_estimators': 200, 'min_samples_split': 5, 'min_samples_leaf': 2, 'max_depth': None, 'criterion': 'log_loss'}
Accuracy for Best Random Forest Classifier: 88.71 %
F1 score of Test set: 0.4680851063829788
```

3. ***AdaBoost Classifier*** - This model was the best performing model among all models when it was run in its default setting (without any tuning) having an accuracy score of 87.58% and a F1 Score of 0.513

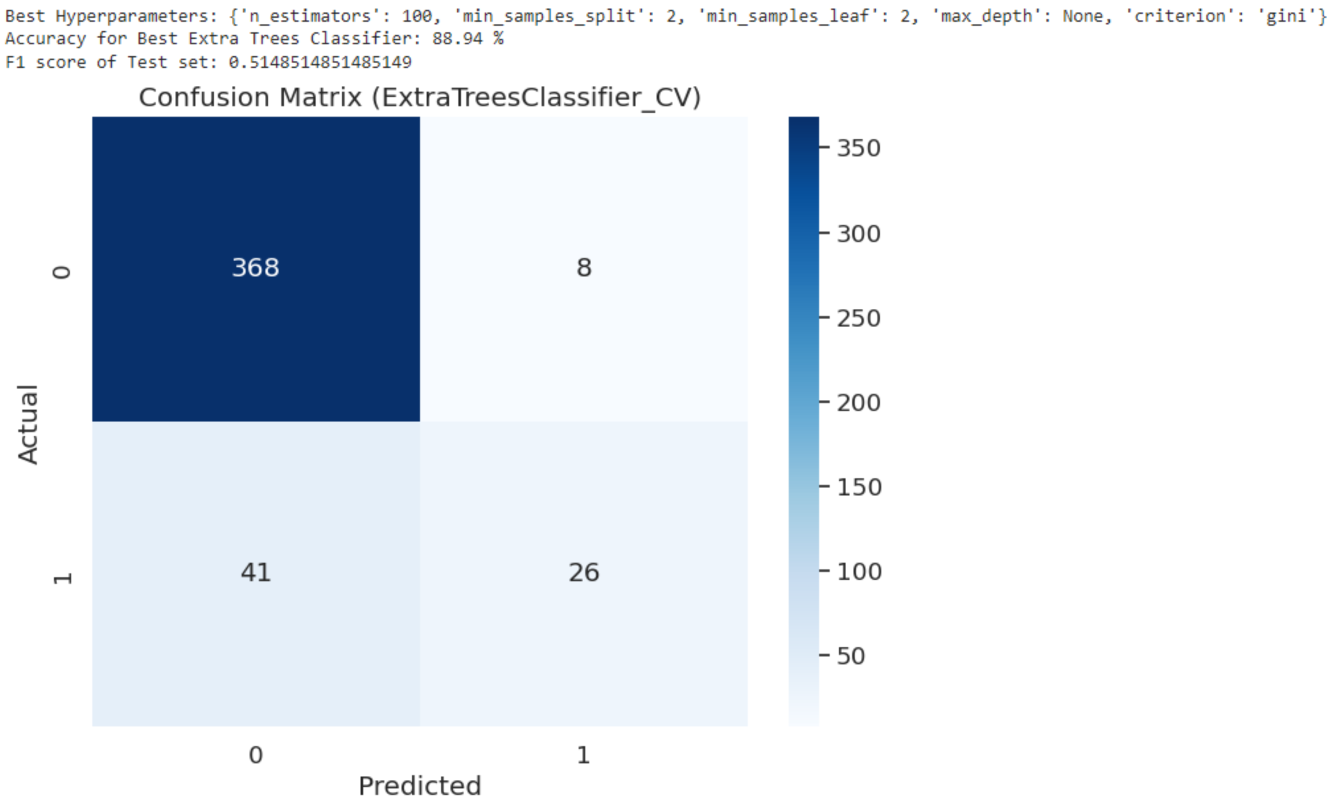
   Best AdaBoost Model - *Without Hyperparameter tuning*



```
AdaBoost Accuracy: 87.58 %
F1 score of Test set: 0.5132743362831859
```

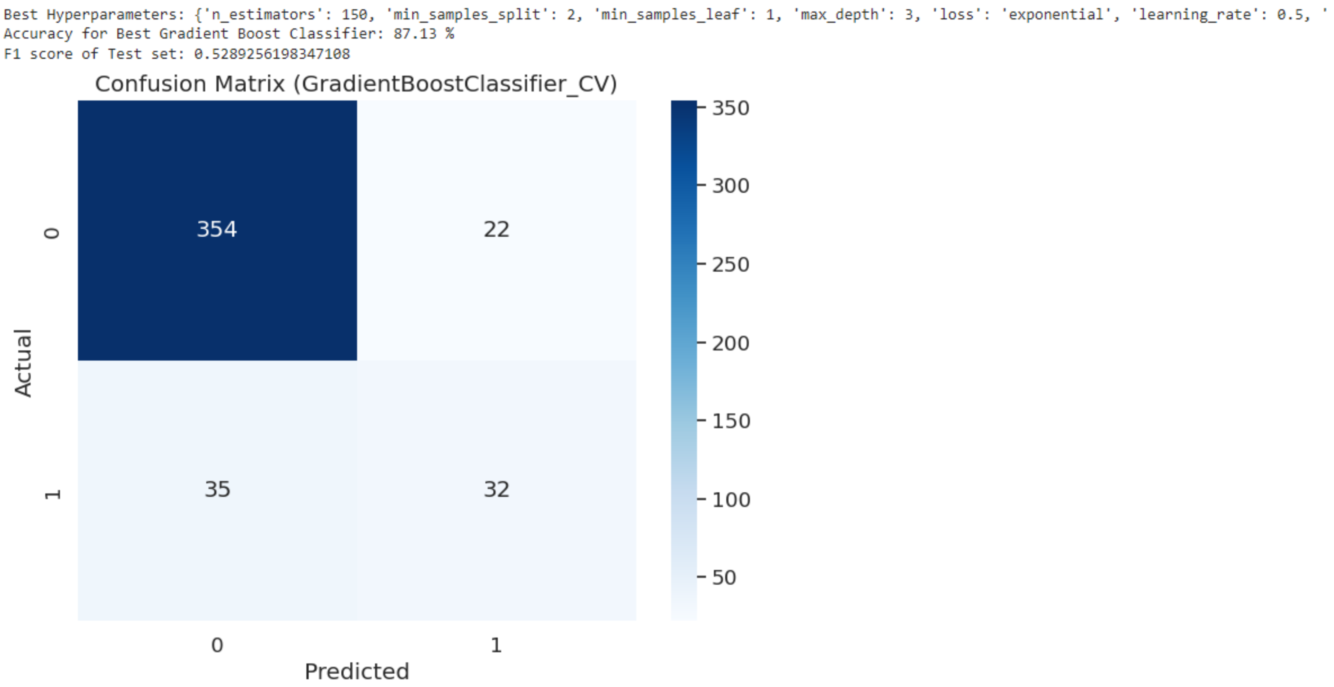
4. ***Extra Trees Classifier*** - This model displayed splendid performance with hyperparameter tuning achieving an accuracy score of 88.94% coupled with a F1 score of 0.515

   Best Extra Trees Model - *With Hyperparameter tuning*



```
Best Hyperparameters: {'n_estimators': 100, 'min_samples_split': 2, 'min_samples_leaf': 2, 'max_depth': None, 'criterion': 'gini'}
Accuracy for Best Extra Trees Classifier: 88.94 %
F1 score of Test set: 0.5148514851485149
```

5. ***Gradient Boosting Classifier*** - This algorithm displayed great performance both with and without hyperparameter tuning and had the best performance metrics among all the models when tuning was performed, with an accuracy score of 87.13% and an F1 score of 0.529



```
Best Hyperparameters: {'n_estimators': 150, 'min_samples_split': 2, 'min_samples_leaf': 1, 'max_depth': 3, 'loss': 'exponential', 'learning_rate': 0.5, '
Accuracy for Best Gradient Boost Classifier: 87.13 %
F1 score of Test set: 0.5289256198347108
```

Modeling Results

The best model is Gradient Boosting Classifier as it has the highest F1 scores among all models. Since this is an imbalanced dataset, it made sense to leverage F1 scores instead of accuracy scores for evaluating model performance.

This model can now be used to predict who should be targeted for campaigns.

Conclusion

In conclusion, customers' personality traits have a significant impact on their purchase behavior. Our analysis has revealed some common personality traits among our customers, allowing us to personalize our offer and communication . We have identified distinct customer groups with unique purchasing behavior, and we now have a clear understanding of our best customers.The integration of customer personality data, past purchase behavior and campaign response metrics has shown promising potential for predicting how customers will respond to the upcoming campaigns. Leveraging this predictive model, we can optimize our marketing efforts as well as maximizing the return on investment for each campaign.