Optimizing Marketing Campaigns: Data-Driven Strategies for Enhanced Profitability

Stella Cervini

July 16, 2024

1 Introduction

In the fast-paced world of direct marketing, predicting customer behavior accurately is key to boosting campaign profitability. This presentation delves into the analysis of a marketing campaign for a new gadget. The aim is to develop a predictive model to identify the most responsive customers with the goal of helping to optimizing the campaign efficiency, reduce costs, and increase revenue.

2 Explorative analysis

The dataset used in this project consists of information from a marketing campaign aimed at promoting a new gadget. It includes data for 2,240 customers and encompasses various attributes that can influence purchasing decisions. The analysis was sectioned in three main parts, each of which focuses on a different component of the data:

- Customer's personal information
- Information about Customer behaviour
- Information about marketing results

2.1 Customer's personal information

The analysis of customers' personal information focuses on understanding the various demographic and socio-economic attributes of the customers participating the study. The variables taken into account reefer to age, income, education level, marital status, and the number of children at home (`Kidhome` and `Teenhome`).

2.1.1 Marital status and education

The first variables that we are going to study are `Marital_Status` and `Educaiton`, representing the customer's marital status and customer's level of education respectively.
To visualize the data, we use bar charts, as shown in Figure 1.
The first chart shows the number of people in each education and marital status category.

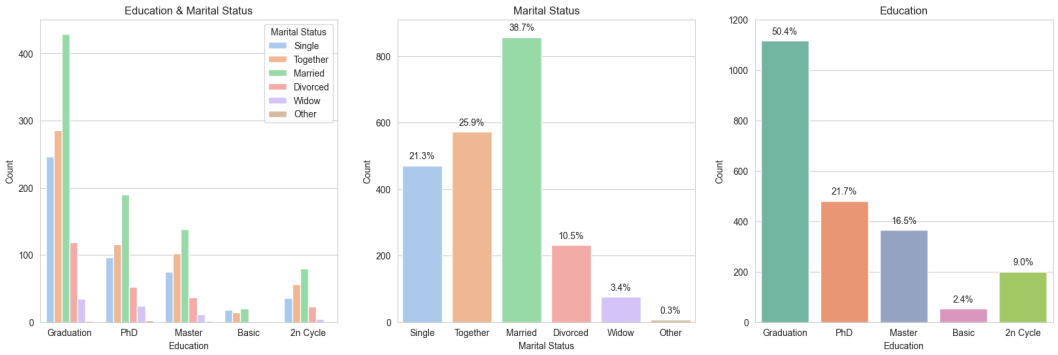

Figure 1: Marital status and education

The second chart shows the percentage of people in each marital status category, and the third chart shows the percentage of people in each education category.

In the dataset, the majority of people fall in the education categories Graduation, followed by PhD and Master, while Married, Single and Together are the most common values in `Marital_Status`, with more that 85% of customers falling in one of these categories (more than a third of the participants are Married).
There is a positive correlation between education and marital status and the plot suggests that education and marital status are strongly related. People with higher levels of education are more likely to be married and less likely to be single.

2.1.2 Children

Focusing on the number of children in the households, two variables can be taken into account: (`Kidhome` and `Teenhome`). They represent the number of small children and teenagers in customer's household. A third variable was created, `Children`, to highlight the overall presence of children in the families. The new variable is the sum of the two preexisting ones.

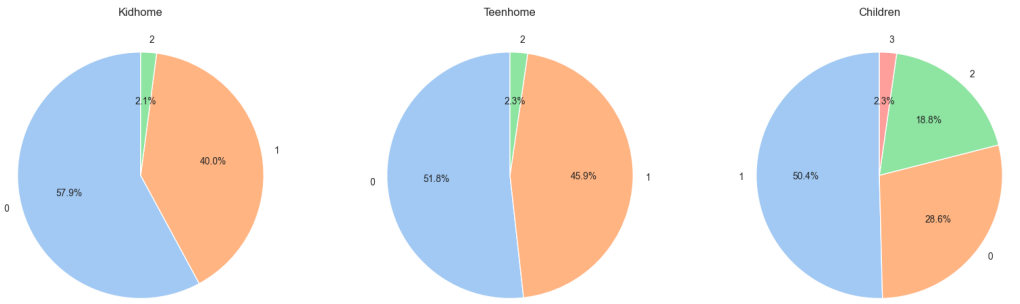

Figure 2: Percentage of small children, teenagers and overall children in customer's household.

From the piecharts in Figure 2, we can observe that the distribution of the number of children in a household is relatively stable across different groups of households. Infact, small children and teenager are distributed in a similar way and in both cases more than a

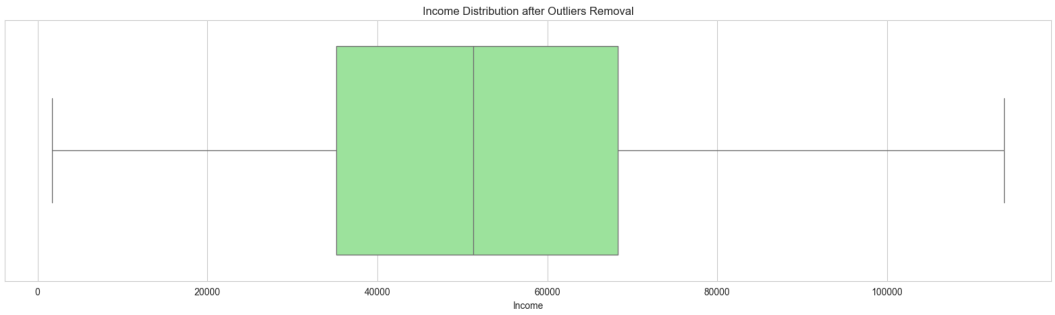

Figure 3: Income Distribution after Outliers Removal

half of the households do not present children. It is also extremely rare for a household to have 3 children.

2.1.3 Income

Our analysis of income distribution reveals a compelling picture of financial disparity (Figure 3). The box plot shows that the median income is around 50,000, and the interquartile range (IQR) is around 20,000. The whiskers extend to the minimum and maximum values, which are around 20,000 and 80,000, respectively.
The distribution of income is right-skewed, meaning that there are more people with lower incomes than higher incomes (note that the plot shows income distribution after outliers removal - the very high incomes have been removed).
The IQR is relatively small, which means that most of the income is concentrated in a narrow range. The median income provides a more accurate representation of the typical income level compared to the mean, which is likely inflated by a small number of high-income earners.
The majority of individuals fall within a relatively narrow income range, suggesting a concentrated income distribution. However, despite the overall concentration, there's a notable presence of both lower and higher income outliers, indicating income polarization within the population.

2.1.4 Age

Regarding customers' age, the provided histogram Figure 4 visualizes the distribution of ages after outliers have been removed from the dataset, ensuring that the observed distribution accurately reflects the central tendency and variability of the data without being influenced by extreme values. The x-axis represents age, ranging from approximately 30 to 80 years, and the y-axis indicates the count of individuals within each age bin. The overlaying curve represents the kernel density estimation (KDE), which smooths the histogram and provides a continuous approximation of the data distribution.

The age distribution appears to follow a roughly normal distribution, with a peak around the 50-55 age range. This suggests a relatively balanced representation of ages within the

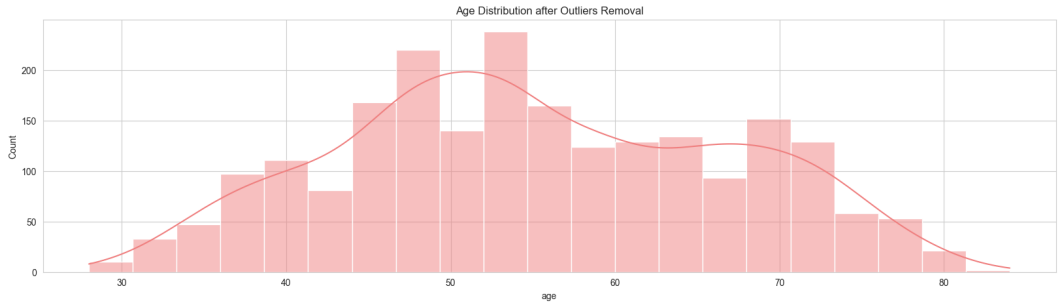

Figure 4: Age distribution

dataset.
The majority of individuals in the dataset are between 40 and 60 years old, indicating a potential focus on a specific age group.
Note that `age` is computed from the variable `Year_Birth` (Customer's year of birth).

2.1.5 Overview

The next plot (Figure 5) wants to provide a synthesis of the information seen until this point. It visualizes the distribution of age across different education and marital status categories (`Income` here is not considered for two main reasons: scale incompatibility and we can assume that higher salaries are associated with higher level of education and higher age).

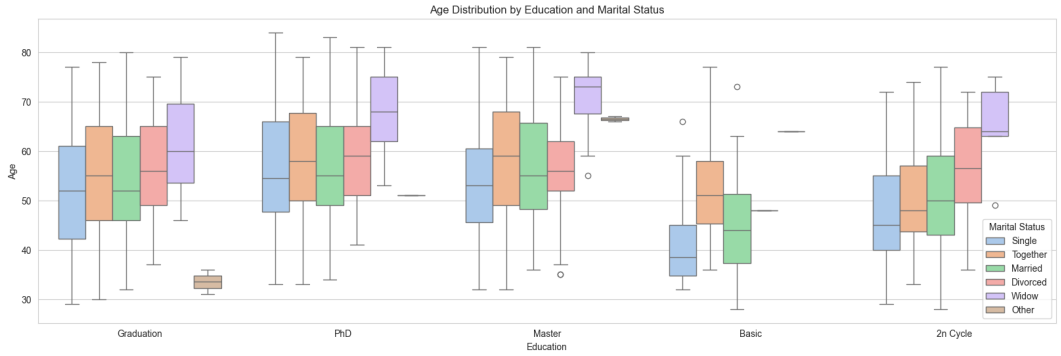

Figure 5: Age Distribution (without outliers) partitioned Education and Marital Status.

The x-axis represents education levels, anging from "Graduation" to "2n Cycle", while the y-axis indicates age. Each box plot represents a specific marital status category, color-coded for differentiation.
There is a trend of increasing median age with higher education levels, that si individuals with higher education tend to be older, on average.
Marital status also impacts age distribution. Generally, individuals with higher marital status categories (e.g., "Married", "Divorced", "Widow") tend to be older compared to "Single" or "Together". This is supported by the higher median age in these categories.

The interquartile range (IQR), represented by the box height, indicates the spread of ages within each group. There is considerable variation in age within each education and marital status category.
The presence of outliers, represented by individual data points beyond the whiskers, suggests potential anomalies or extreme cases within certain education and marital status combinations. For example, there are a few individuals with exceptionally high ages in the "Divorced" and "Widow" categories.

2.2 Analysis of Customer Purchase Behavior

This section focuses on on spending across various product categories, recency of purchases and place of purchase (online or in physical shops). Another variable taken into account is `client_time_years`, representing for how many years a person has been customer. A `sage`, this variable was not initially in the dataset but was computer from `Dt_Customer`, representing the date of customer's enrolment with the company.
This analysis helps identify trends and preferences, make us able to understand purchase behavior, thus enabling more targeted and effective campaigns, enhancing customer engagement and profitability.

2.3 Purchase recency and client time

The provided plot presents a comparative analysis of two key customer metrics: Purchase Recency and Client Time. Both metrics are visualized using box plots, as in Figure 6, allowing for a direct comparison of their distributions.

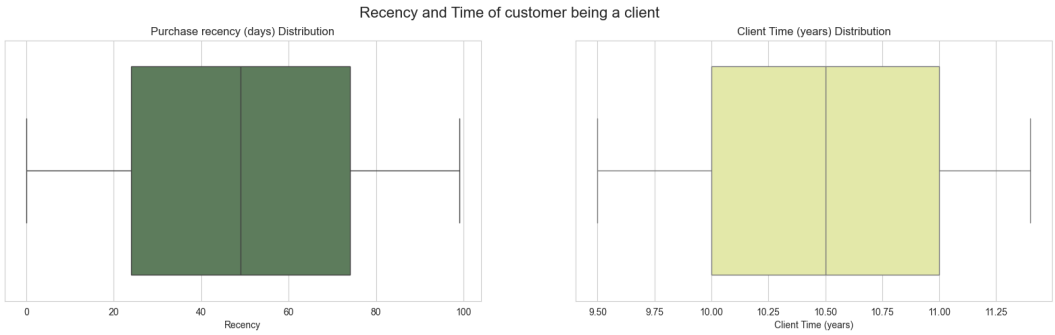

Figure 6: Purchase recency and client time

The distribution of purchase recency (in days) is concentrated between 0 and 80 days. The median purchase recency lies around the 40-day mark, suggesting a moderate frequency of recent purchases among the customer base. The relatively narrow interquartile range indicates a consistent purchase pattern within this timeframe.
The distribution of client time (in years) is centered around the 10-year mark, with a tight distribution. This suggests a relatively stable customer base with a substantial portion having been clients for approximately 10 years. The narrow interquartile range indicates a consistent customer tenure.

Both purchase recency and client time exhibit concentrated distributions with minimal outliers. This suggests a relatively homogeneous customer base in terms of purchase behavior and customer tenure.

2.3.1 Amount of money spent on different products

The data regarding the amount spent on different products types can help us identify with are the most loved products on the catalog.
The image in Figure 7 shows a violin plot of the amount of money spent on different products. The x-axis shows the different products, and the y-axis shows the amount of money spent. The violin plot shows the distribution of the data, with the width of the violin representing the density of the data at that point. The white dot in the middle of each violin is the median, and the black line is the interquartile range (IQR).

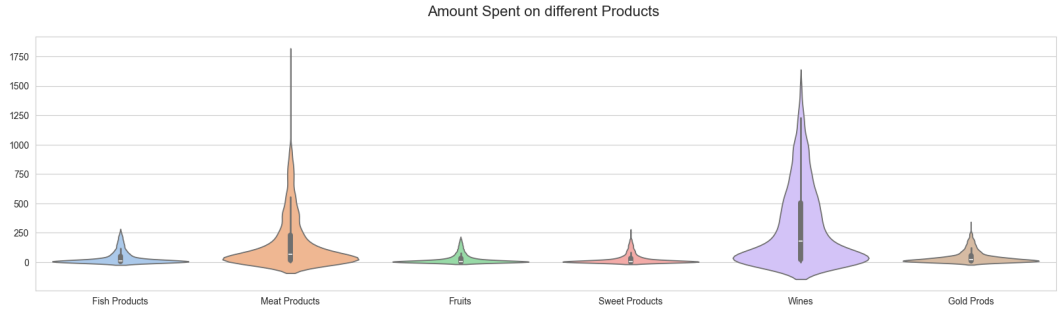

Figure 7: Amount of money spent on different products

The most money is spent on wine, followed by meat products, fish products, fruits, sweet products, and gold products. The distribution of the amount of money spent on wine is the widest, which means that there is a lot of variation in the amount of money that people spend on wine. The distribution of the amount of money spent on wine is also the widest, which means that there is a lot of variation in the amount of money that people spend on wine.
The distribution of the amount of money spent on meat products and fish products is similar, with a smaller number of people spending a lot of money on these products, indicating that customers spending on this product is very subjective. This could also imply that meat and fish products has a wider range of price themselves.
The distribution of the amount of money spent on fruits and sweet products is also similar, with a smaller number of people spending a lot of money on these products. This could suggest that prices are not very high but there are a few people that consumes a lot of this kinds of products.
The distribution of the amount of money spent on gold products is the narrowest, which means that there is less variation in the amount of money that people spend on gold products. We can also assume that spending on gold products is not very common and not every customer is willing to make such purchase.

2.3.2 Purchase Variables - place and methods for purchase

This section presents a comparison of the distribution of five different purchase variables:

- Number of Deals Purchased
- Number of Web Purchases
- Number of Catalog Purchases
- Number of Store Purchases
- Number of Web Visits per Month

These variables represents where customers do their shopping (on the Web, including how many times that visit the website, or in physical stores) and which method they use, meaning, for example, if they prefer to wait for deals or buy directly from the catalog.

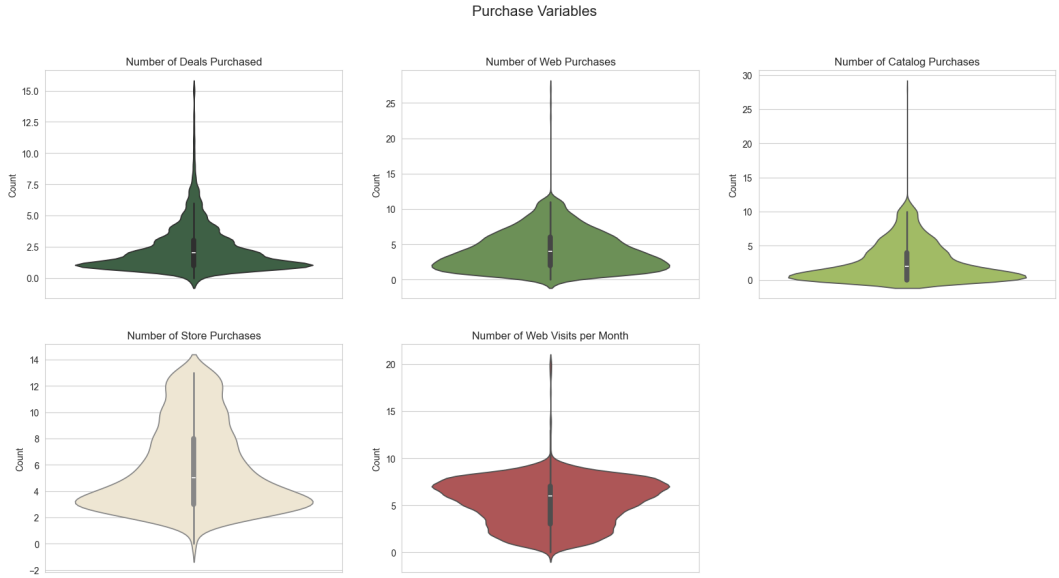

Figure 8: Purchase Variables

As for the analysis of amount of money spent on each type of product, each variable here is visualized using a violin plot (Figure 8), which displays the distribution of the data points. The width of the violin represents the density of the data at that point, while the white dot indicates the median and the black line represents the interquartile range (IQR). he distributions of the purchase variables exhibit varying shapes. Some, like the number of web purchases and catalog purchases, appear to be relatively symmetric, while others, such as the number of store purchases, show a more skewed distribution.
The median values provide insights into the typical number of purchases or visits for each variable. For instance, the median number of web purchases is around 15, suggesting that a significant portion of customers make around this number of web purchases.
The width of the violin plots indicates the variability in the data. A wider violin suggests a larger spread of values, while a narrower violin indicates less variability. In this case,

the number of web visits per month shows a wider distribution compared to the number of catalog purchases, implying more variation in website visits among customers.
The presence of outliers, represented by individual data points beyond the violin body, can be observed in some variables. These outliers indicate extreme values that deviate significantly from the main distribution.

2.3.3 Amount spent in physical stores and online

Here the analysis focus on visualizing the amount spent on different kind of product in physical stores and online. The insights retrieved from the images could help us identify spending habits, highlighting if there are some products that customers prefer to buy in person or online.

Starting from physical locations, the provided stacked bar chart in Figure 9 visualizes the mean amount spent on different product categories within physical stores, categorized by the number of purchases made. The x-axis represents the number of purchases made in person in physical stores, while the y-axis represents the mean amount spent. Each product category (fish products, meat products, fruits, sweet products, wines, and gold products) is represented by a different color, stacked on top of each other for each number of purchases.

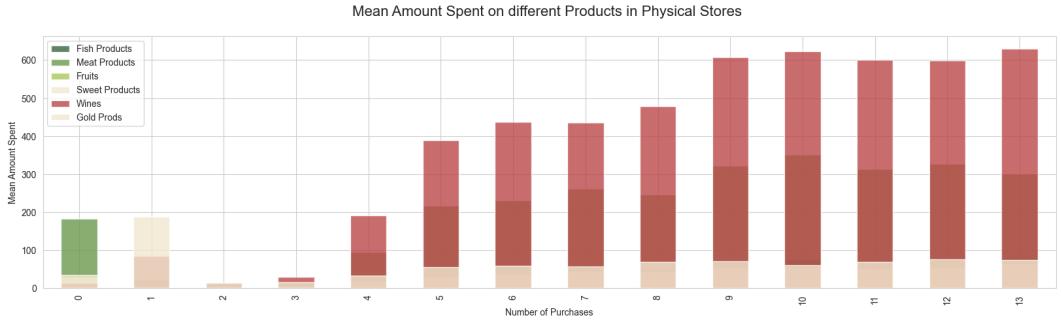

Figure 9: Mean amount spent on different products in physical stores, stacked by product type.

The overall trend shows that as the number of purchases increases, the mean amount spent on most product categories also increases. This suggests that customers who make more frequent purchases tend to spend more per purchase on average.
Wines and meat products consistently account for the largest proportion of spending across different purchase frequencies. This indicates that these two product categories are the primary drivers of revenue in physical stores.
While the mean amount spent on fish products increases with purchase frequency, the growth is less pronounced compared to other categories.
Fruits and sweet products categories show a relatively consistent spending pattern across different purchase frequencies, with moderate growth while the spending on gold products is generally lower compared to other categories and exhibits less variation with the number of purchases, suggesting that customers may prefer to buy gold products online.
There is considerable variation in spending patterns across different product categories and

purchase frequencies, indicating diverse customer preferences and behaviors.

Let us now focus on online purchases. The stacked bar plot shown in Figure 10 represents similar data as the previous one but this time on the x-axis we have the number of purchases made online.

Figure 10: Mean amount spent on different products in the web store, stacked by product type.

As seen previously, the trend shows that as the number of purchases increases, the mean amount spent on most product categories also increases, observation from which we can draw the same conclusion as before (customers who make more frequent purchases online tend to spend more per purchase on average).
Wines and meat products are still accountable for the largest proportion of spending. A similar behaviour to the latter plot can be seen also for fish fruit products.
However, we can observe that the spending on sweets is much higher here towards the right than in the previous plot, meaning that customers prefer to buy sweets online.

2.4 Information about marketing

Now we analyzed marketing campaign data, including response rates and complaints. We will also address the cost and revenue per campaign.

2.4.1 Response rate per campaign

The image in Figure 11 shows a bar chart of the number of accepted and not accepted campaigns. The x-axis shows the different campaigns, and the y-axis shows the number of people. The red bars represent the number of people who did not accept the campaign, and the green bars represent the number of people who accepted the campaign.

We can observe that campaigns 3, 4, and 5 were the most successful ones, with campaign 3 and 4 reaching the 7.4% of response rate. The less popular campaign was the number 2, with the lowest response rate (1.45%). The reason for this discrepancies could be be that some campaigns are more relevant to people's needs or interests than others.

2.5 Complaints and overall response rate

Below (Figure 12) are visualized the he distribution of complaints and responses.

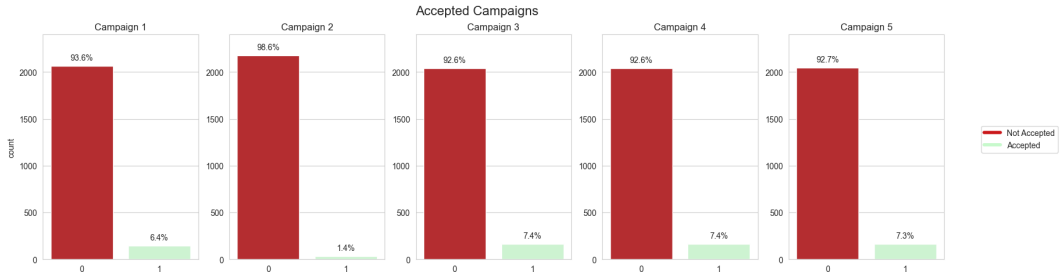

Figure 11: Accepted Campaigns rate.

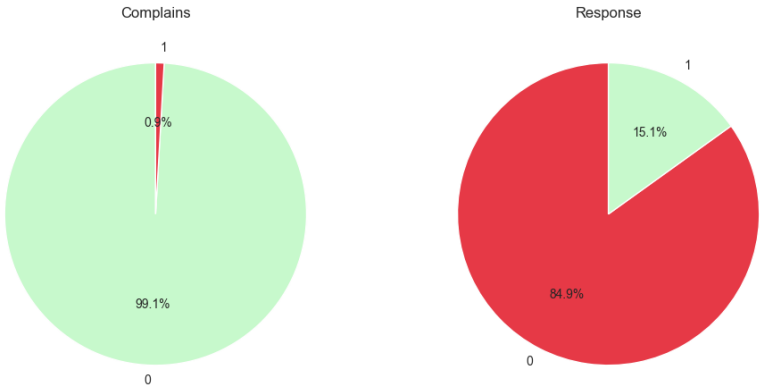

Figure 12: Complains and responses percentages.

The first chart shows that 99.1% of the customers did not complain, while 0.9% of the customers complained, indicating that the majority of customers are satisfied with the company and do not need to complain or respond to the campaign. However, the company should still investigate the reasons why some customers complained and try to improve its products or services.
On the other hand one could argue that campaigns could be more successful. Infact, the second chart shows that 84.9% of the customers did not respond to the campaign, while 15.1% of the customers responded to the campaign, meaning that overall a larger percentage of customers responded to the campaign than complained. However, the company should also try to increase the response rate to its campaigns.

2.5.1 Cost and revenue per campaign

The average expenditure incurred for each campaign and the average income generated by each campaign can be seen in Table 1.

Cost per campaign	3.0
Revenue per campaign	11

Table 1: Campaign Metrics

2.6 Correlation

To explore the relationships among various customer attributes and their impact on purchasing behavior correlation analysis was conducted. A visual representation of the correlation coefficients between different variables in a dataset is shown in Figure 13.

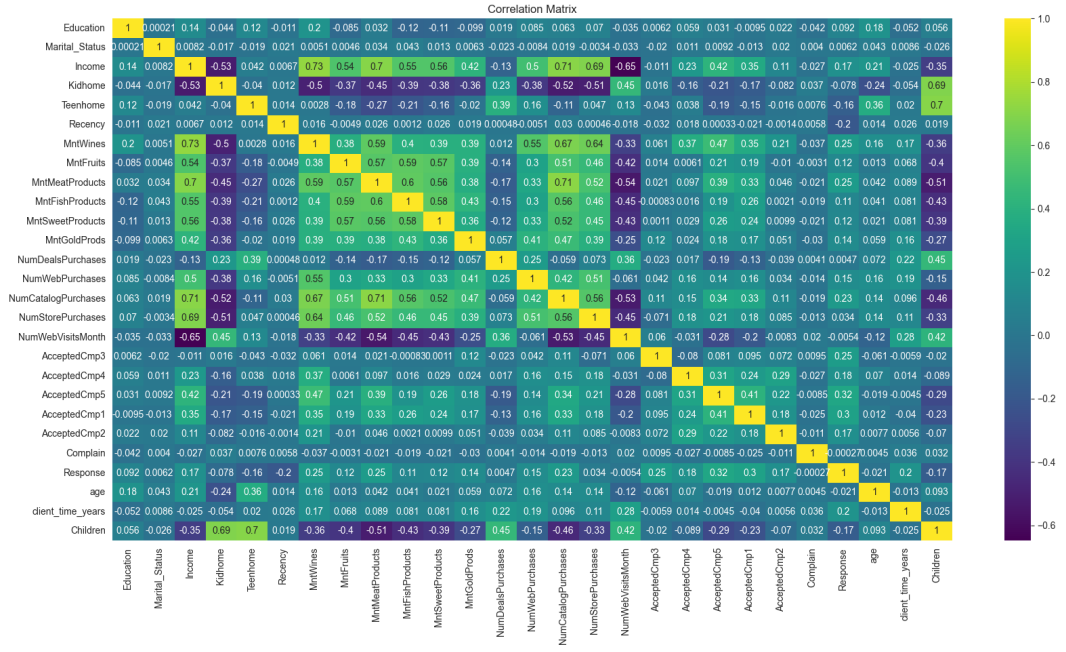

Figure 13: Correlation matrix

Several variables exhibit strong positive correlations, such as:

- `Income` and `Education`
- `Kidhome` and `Teenhome` (indicating households with kids are more likely to have teenagers)
- `MntWines` and `MntMeatProducts` (suggesting overall spending patterns)
- `NumDealsPurchases`, `NumWebPurchases`, `NumCatalogPurchases`, and `NumStorePurchases` (indicating overall purchase activity)

On the other hand, some notable negative correlations include:

- `Education` and `Response` (indicating individuals with higher education might be less likely to respond to campaigns)
- `Kidhome` and `Income` (suggesting families with children might have lower income)
- `Teenhome` and `Income` (as above, suggesting families with teens might have lower income)
- `Recency` and `MntWines`, `MntMeatProducts`, `MntFruits`, `MntSweetProducts`, `MntGoldProds` (indicating recent customers tend to spend more)

Most of the correlation coefficients are relatively small, suggesting that the relationships between many variables are weak or non-existent.
The variables with the highest positive correlations are:

Variable 1	Variable 2	Correlation
Kidhome	Teenhome	0.7
MntWines	MntMeatProducts	0.59
NumDealsPurchases	NumWebPurchases, NumCatalogPurchases, NumStorePurchases	around 0.5

Table 2: High Correlations

These variables exhibit a strong linear relationship, meaning that an increase in one variable is associated with an increase in the other.

3 Causal impact of the variable Kidhome on the variable Response

We investigated the potential causal impact of the variable `Kidhome` (number of small children in a customer's household) on the `Response` to the marketing campaign. By employing Logistic Regression and Propensity Score Matching (PSM), we aim to determine whether the presence of small children influences a customer's likelihood to respond, providing insights for more targeted and effective marketing strategies.

The first approach in verifying if there exist a causal impact between the two variables is to plot them (Figure 14).

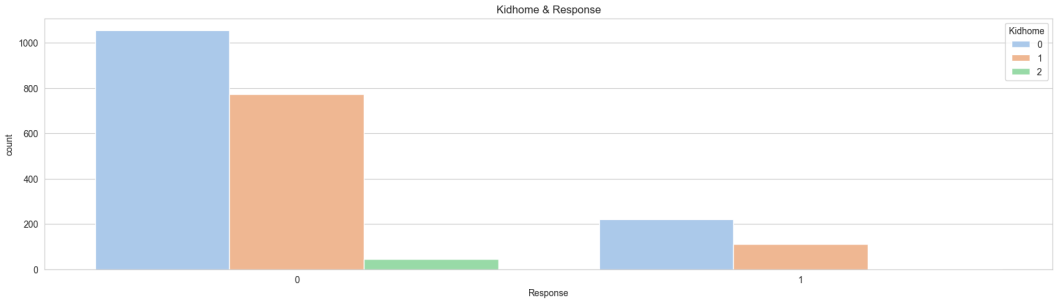

Figure 14: Bar chart of the number of customers who responded to a campaign, grouped by the number of children in their household.

From here we can draw the first conclusions:

1. the majority of customers who did not respond to the campaign had no children.
2. The majority of customers who responded to the campaign had one child.
3. There is a positive correlation between the number of children and the likelihood of responding to the campaign. In other words, customers with more children are more likely to respond to the campaign.

Overall, the plot shows that the number of children in a household seems to be a good predictor of whether or not a customer will respond to a campaign.

As introduced above, a Logistic Regression model was used to assess causal impact. The results are shown in Table 3.

	Logit Regression					
Variable	coef	std err	z	P $>\|z\|$	[0.025	0.975]
const	-1.5536	0.073	-21.237	0.000	-1.697	-1.410
Kidhome	-0.4332	0.118	-3.662	0.000	-0.665	-0.201

P-values of the model:	const	4.309102e-100
	Kidhome	2.498751e-04
Odds Ratios:	const	0.211495
	Kidhome	0.648408

Table 3: Logit Regression Results

The provided table presents the results of a logistic regression model where the dependent variable, `Response`, is binary (likely 0 or 1, indicating whether a customer responded to a campaign). The independent variable is `Kidhome`, representing the number of children in the household.
The intercept of -1.5536 represents the log odds of a customer responding to the campaign when `Kidhome` is 0 (i.e., no children).
The coefficient of -0.4332 indicates that for each additional child in the household, the log odds of a customer responding to the campaign decrease by 0.4332, holding other factors constant.
Both the intercept and `Kidhome`‘ coefficients have p-values less than 0.05, indicating that they are statistically significant at the 5% level. This means that both variables have a significant impact on the likelihood of a customer responding to the campaign.
The odds ratio for `Kidhome` is 0.648408. This means that, for each additional child, the odds of a customer responding to the campaign decrease by approximately 35.16% (calculated as 1 - 0.648408).
Overall, the logistic regression model suggests that having children in the household is negatively associated with the likelihood of a customer responding to the campaign. As the number of children increases, the odds of a customer responding decrease. This information could be valuable for targeted marketing strategies. For instance, the company might consider focusing marketing efforts on households with fewer children.
It is still relevant to notice that additional factors might influence the response to the campaign, and a more complex model might be necessary for a comprehensive analysis (we will use Propensity Score Matching (PSM) for this purpose).

These conclusions can also be visualized in the two following plots (Figure 15 and Figure 16.

The provided bar chart visualizes the relationship between the number of children in a

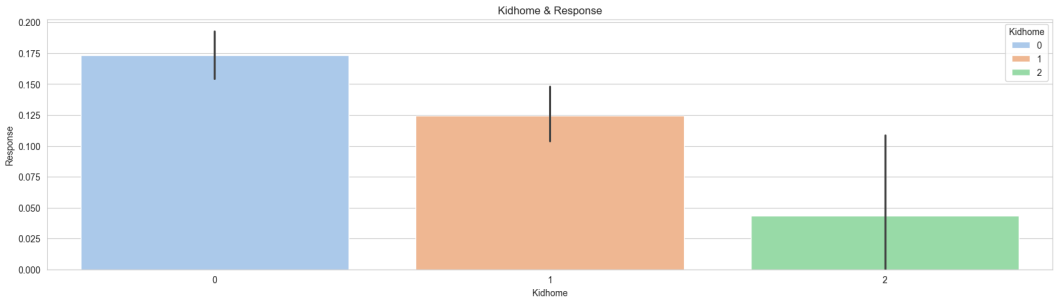

Figure 15: Kidhome and Response barplot

household (`Kidhome`) and the response rate to a campaign. Error bars represent the uncertainty or variability in the response rate for each `Kidhome` category.
The response rate to the campaign decreases as the number of children in a household increases. Households with no children have the highest response rate, followed by households with one child, and then households with two children.
The error bars suggest that the difference in response rates between households with no children and households with one or two children is statistically significant. However, the difference between households with one and two children might not be statistically significant due to overlapping error bars.

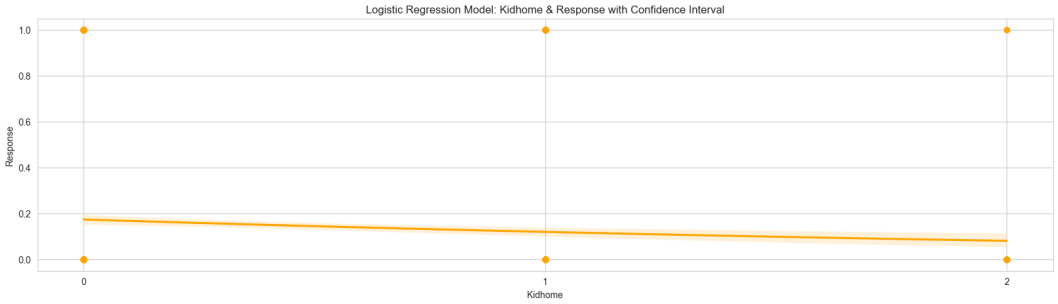

Figure 16: Logistic regression: Kidhome and Response

The plot shows that as the `Kidhome` variable increases, the probability of the response variable being 1 (accepted campaign) decreases.
The curve represents the predicted probability of the response variable based on the logistic regression model. It shows the relationship between the `Kidhome` variable and the probability of the response variable being 1 (accepted campaign). The line is based on the coefficients of the logistic regression model and represents the estimated probability of the response variable for different values of the `Kidhome` variable.
This is consistent with the odds ratio of the `Kidhome` variable in the logistic regression model, which indicated a decrease in the odds of the response variable with an increase in the `Kidhome` variable.

The propensity score is a statistical concept used primarily in observational studies to estimate the effect of a treatment, intervention, or exposure on an outcome by accounting

for the covariates that predict receiving the treatment. It is defined as the probability of assignment to a particular treatment given a set of observed covariates.
In this example the propensity score represents the probability of receiving the treatment (`Kidhome`) based on the covariates (`Income`, `age`, `Marital_Status`).
It is calculated using a logistic regression model that predicts the likelihood of receiving the treatment based on the covariates. The propensity score can be used to balance the treatment and control groups in observational studies by matching or weighting the samples based on the estimated probability of receiving the treatment. Once estimated, propensity scores can be used in several ways. In this scenario, we used to for confounding, including matching. Matching means that individuals in the treatment group are matched with individuals in the control group who have similar propensity scores.
After matching, the Average Treatment Effect (ATT)is estimated by comparing the outcomes (e.g., response variable) between the treated and matched control units across the entire sample.
Below in the table is reported the ATT result obtained.

ATT	0.0635

Table 4: Average Treatment Effect on the Treated (ATT)

The result indicates that, on average, the treatment increases the outcome by approximately 0.0635 units for those units that received the treatment, compared to what their outcome would have been had they not received the treatment. In this context, it means that having children at home (the `Kidhome` variable being the treatment in this case) is associated with an average increase of 0.0635 units in the response variable, after matching based on propensity scores and controlling for other covariates.
This ATT value suggests that for the treated group — individuals with a `Kidhome` value of `1` - the presence of children at home has a positive effect on the response variable, which could be any metric of interest in the context of a marketing campaign, such as purchase amount, frequency of purchase, or likelihood of responding to a campaign. The effect is measured after accounting for other relevant factors that could influence the response variable, providing an estimate of the causal impact of having children at home on the outcome of interest.

4 Market segmentation with SOMs

Self-Organizing Maps (SOMs) are a type of unsupervised learning algorithm used for dimensionality reduction and data visualization, particularly to visualize high-dimensional data in lower-dimensional (usually 2D) spaces. Here, they were used to propose a market segmentation.
The `MiniSom` function was use. It refers to the constructor of the MiniSom class from the MiniSom library, a minimalistic implementation of the Self-Organizing Maps (SOMs). Below (Table 5) are reported the parameters used to create the SOM.

Parameter	Value
Map Height	16
Map Width	16
Sigma	2
Learning Rate	0.7
Number of Features	26
Number of Samples	2205

Table 5: SOM Parameters

By visualizing the distance map created by the algorithm, we can gain a deep understanding of the distribution of data points across clusters.
Figure 17 shows the Self-Organizing Map (SOM) distance map. The x-axis and y-axis represent the dimensions of the SOM grid, and the color of each cell represents the distance between the corresponding node and its nearest neighbor. Darker colors represent shorter distances.

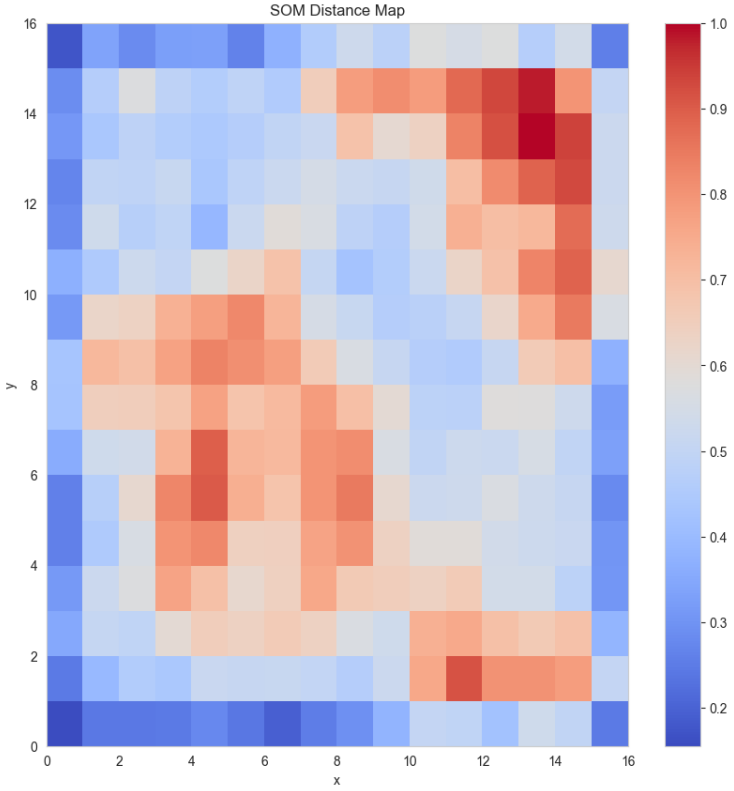

Figure 17: SOM distance map

The SOM is well-organized, with clear clusters of nodes that are close to each other. This suggests that the SOM has successfully captured the underlying structure of the data. The clusters of nodes are not uniformly distributed across the SOM grid, but rather form distinct regions. This suggests that the data has some underlying heterogeneity.
The distance between the nodes in the clusters is relatively small, while the distance between

the clusters is relatively large. This suggests that the clusters are well-separated from each other.
Overall, the SOM distance map shows that the SOM has successfully captured the underlying structure of the data and that the data has some underlying heterogeneity.

To make a step forward toward the creation of market segmentation cluster, the K-Means algorithm was used. The main advantange is enhanced clustering interpretation. Infact, SOMs provide a way to visualize high-dimensional data in a two-dimensional map, where similar data points are mapped close to each other, preserving topological properties. However, interpreting the clusters directly from the SOM can sometimes be challenging, especially with complex datasets. Applying K-Means to the SOM's output can help by clearly defining clusters, making it easier to interpret the results.

The K-means algorithm found 3 clusters, since this is the number of `k` that returned the higher silhouette score. The silhouette score is a measure of how well each data point fits into its assigned cluster. A higher silhouette score indicates that the data points are well-clustered.

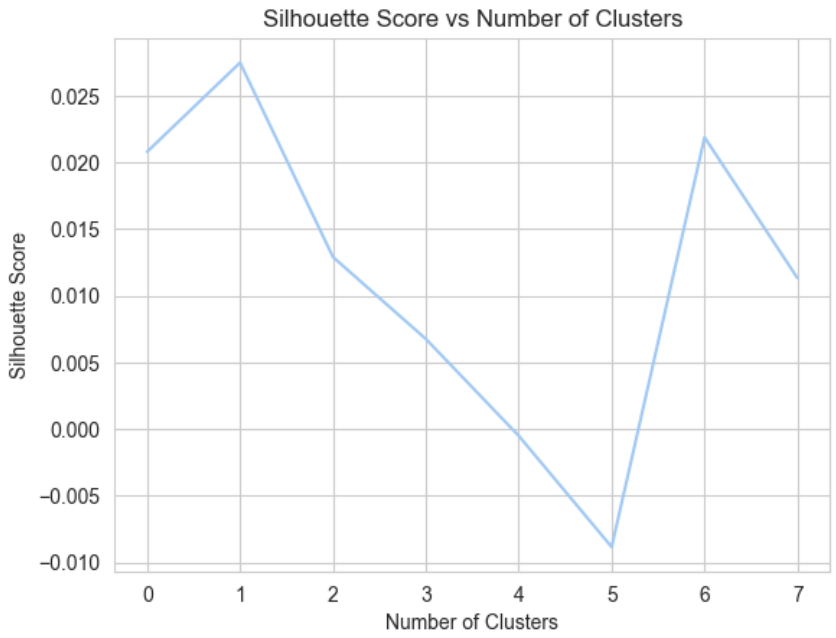

Figure 18: silhouette score

Figure 19 aims to visualize both the SOM distance map and the clusters obtained by applying the K-Means.

The clusters are spatially organized within the SOM grid, indicating that the SOM has effectively captured the relationships between data points.
There is a clear separation between the clusters, as evidenced by the distance between the star markers of different colors. This suggests that the identified clusters are relatively well-defined.

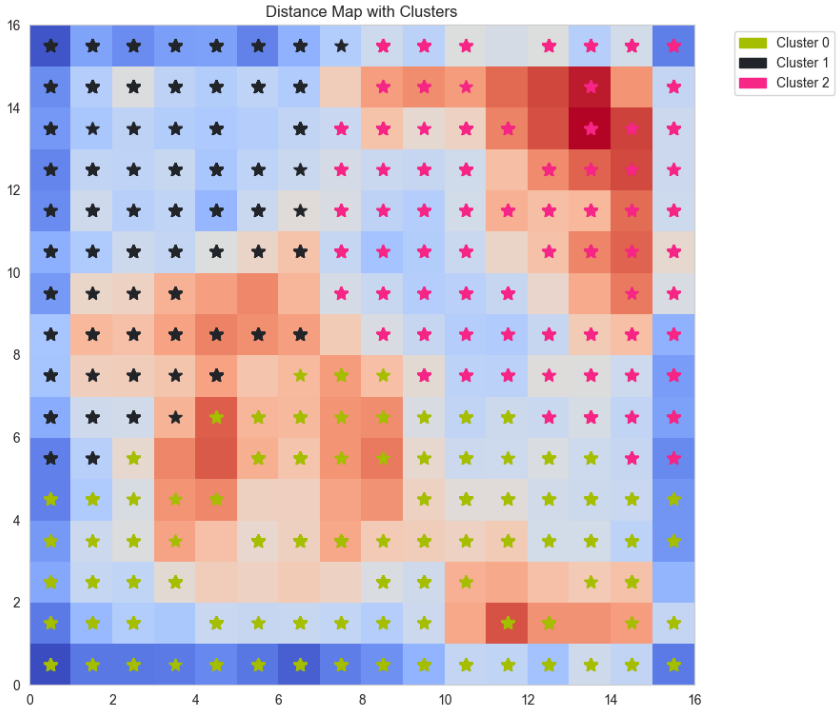

Figure 19: SOM distance Map with Clusters found by K-Means algorithm.

The density of nodes within each cluster varies, with some clusters being more densely populated than others. This could indicate differences in the size or variability of the underlying data groups.
The presence of isolated nodes without clear cluster affiliation might indicate potential outliers or noise in the data.

The last part of this section focuses on cluster size and average response rate. These two metrics, where the first is the number of samples (or data points) assigned to each of the three clusters (0, 1, and 2) and the latter the average response rate within each of the three clusters. Cluster size and average response rate can provide valuable insights into the effectiveness of a clustering algorithm and the characteristics of the resulting clusters.

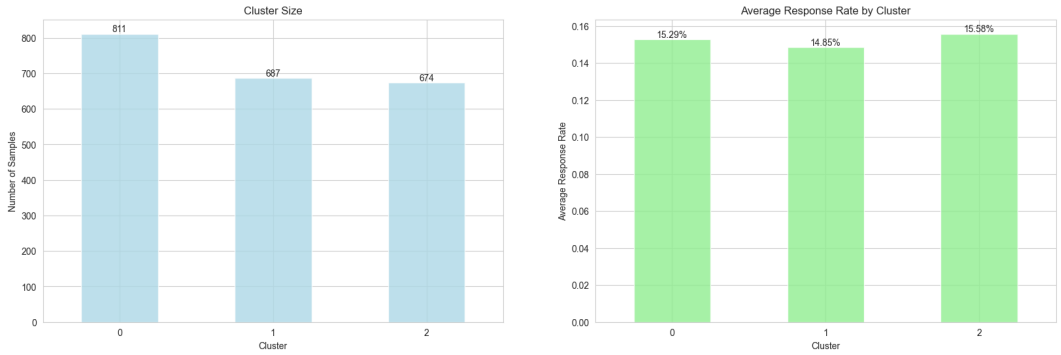

Figure 20: Cluster size and average response rate

Regarding cluster size, cluster 0 is the largest, followed by cluster 1 and cluster 2. There is a noticeable difference in size between the clusters. There is a relatively small difference in the average response rate across the three clusters. The rates are quite similar, ranging between 14.85% and 15.58%. There seems to be no clear correlation between cluster size and average response rate. Larger clusters do not necessarily have higher or lower response rates compared to smaller clusters.
The data suggests that while the clusters have different sizes, they exhibit similar response rates. This indicates that the clustering algorithm might have identified groups of samples with similar characteristics in terms of response rate, but these groups do not significantly differ in terms of their overall size.

5 Comparison of Prediction Models for Response

We compared two different prediction models to forecast the response variable: Logistic Regression and Random Forest. This comparison aims to evaluate the performance of each model in predicting customer responses. Additionally, we analyzed the distribution of predicted probabilities for both `response == 0` and `response == 1` on the test set, providing insights into the accuracy and reliability of each model.

Figure 21 presents a comparative analysis of predicted probabilities generated by the two models. Each plot displays histograms of predicted probabilities for two classes (0 and 1) of the target variable.

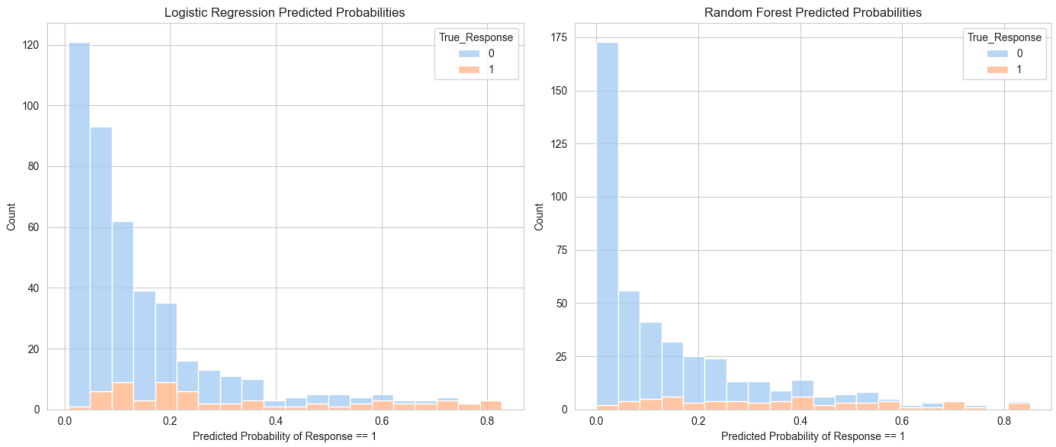

Figure 21: Predicted Probabilities for Logistic Regression and Random Forest.

Both models produce a wide range of predicted probabilities, spanning from nearly 0 to almost 1. This indicates that both models consider a variety of instances with different levels of confidence in their predictions.
There is substantial overlap between the predicted probabilities for classes 0 and 1 in both models. This suggests that the two classes might not be perfectly separable based on the available features.

By analyzing the ROC Curves and Cosnfusion matrices we can draw further conclusions about their performance. ROC Curves compare the performance of both models in terms of True Positive Rate (TPR) against False Positive Rate (FPR) across different classification thresholds, while confusion Matrices Display the number of correct and incorrect predictions made by the two models.

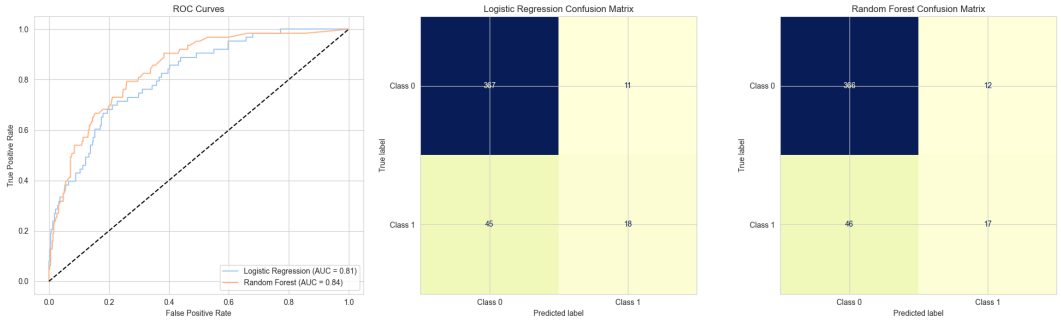

Figure 22: ROC Curves and Confusion Matrices for Logistic Regression and Random Forest.

The ROC curves show that both Logistic Regression and Random Forest achieve good performance, with AUC values of 0.81 and 0.84, respectively. This indicates that the Random Forest model has slightly better discriminative power than Logistic Regression.
The confusion matrices provide a breakdown of correct and incorrect predictions for each class. Both models exhibit relatively high accuracy, with a majority of instances being correctly classified.
The confusion matrices suggest a potential class imbalance, as the number of instances in class 0 is significantly higher than in class 1. This imbalance might affect the model's performance and should be considered during evaluation.
While both models show good performance, the Random Forest model appears to outperform Logistic Regression based on the slightly higher AUC value.

Finally,a complete comparison can be done by observing performance metrics. Based on the table (Table 6), both models exhibit high accuracy, with Logistic Regression slightly edging out Random Forest (0.873 vs. 0.868). This suggests both models are generally good at predicting the correct class. However, F1 Score, precision, and recall metrics reveal a different story. Both models struggle with these metrics, particularly with recall. This indicates that both models have difficulty identifying positive cases correctly.

Model	Accuracy	F1 Score	Precision	Recall
Logistic Reg	0.873	0.391	0.621	0.286
Random Forest	0.868	0.37	0.586	0.27

Table 6: Model Performance Metrics

While both models show high accuracy, their performance on other metrics is suboptimal. This suggests a potential class imbalance problem, where one class (likely the negative class) is significantly overrepresented.

Addressing class imbalance and potentially exploring other evaluation metrics might provide a more comprehensive understanding of the models' performance.

The tables below (Table 7 and Table 8) present the importance scores for two models, derived from feature importance metrics used in model building.

Feature	Importance
ever_accepted_campaigns	0.17
NumWebVisitsMonth	0.13
NumCatalogPurchases	0.13
NumWebPurchases	0.10
AcceptedCmp3	0.09

Table 7: Top 5 Important Features (Logistic Regression)

Feature	Importance
Recency	0.09
Income	0.07
MntWines	0.07
client_time_years	0.07
ever_accepted_campaigns	0.06

Table 8: Top 5 Important Features (Random Forest)

The variable `ever_accepted_campaigns` appears in both lists with relatively high importance scores, suggesting it is a strong predictor for both models. `NumWebVisitsMonth` and `NumCatalogPurchases` are also among the top five features in the first model, indicating their potential predictive power.
The second model highlights different features as important, with `Recency`, `Income`, `MntWines`, and `client_time_years` ranking higher than in the first model. This suggests that the two models might be capturing different aspects of the data or using different algorithms to assess feature importance.
The high importance of `ever_accepted_campaigns` suggests that incorporating features related to past campaign interactions could be beneficial for model performance. The differences in feature importance between the two models indicate that comparing multiple models and their feature importance rankings can provide insights into the underlying data structure and the strengths and weaknesses of different modeling techniques.

6 Profit Curve Comparison of Prediction Models

We compared the two prediction models—Logistic Regression and Random Forest—using profit curves to determine their effectiveness in maximizing campaign profits. By analyzing the profit curves, we aim to identify the optimal percentage of customers to contact in order to achieve the highest possible profitability. This analysis helps in making data-driven decisions for future marketing strategies, ensuring maximum return on investment.

The plot in Figure 23 summarizes the profit curves from the two models. The x-axis represents the threshold for contacting customers, ranging from 0 to 1. The y-axis represents the profit generated, with negative values indicating losses.

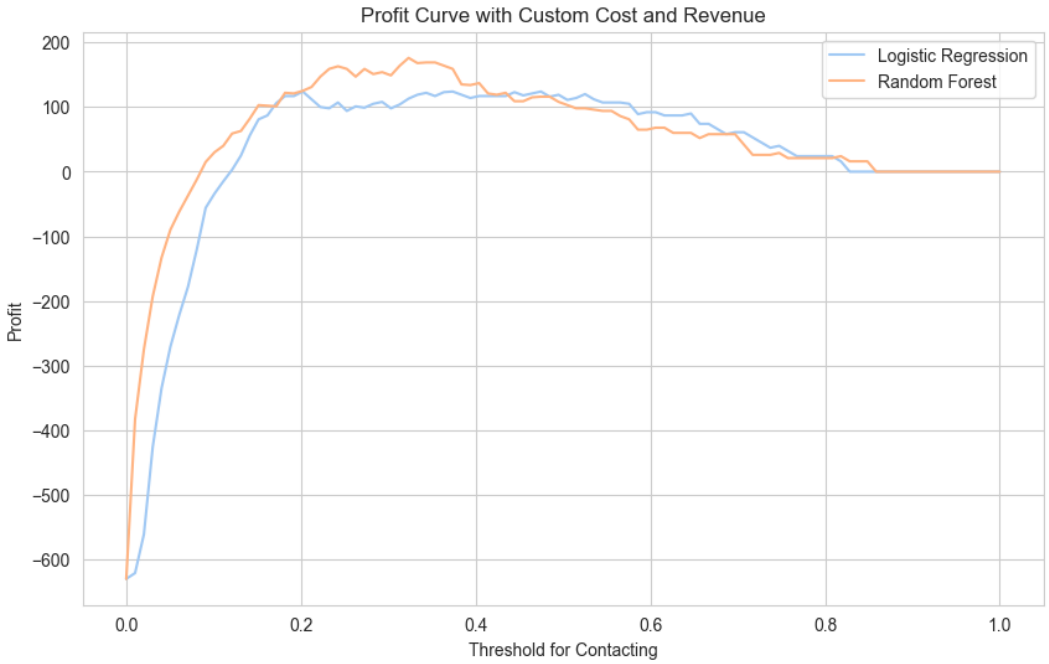

Figure 23: Profit Curve with Cost and Revenue (Logistic Regression and Random Forest)

Both models exhibit a non-linear relationship between the contact threshold and profit. There exists an optimal threshold for each model that maximizes profit.
The Random Forest model appears to outperform Logistic Regression in terms of maximum achievable profit. The peak of the Random Forest curve is higher than that of the Logistic Regression curve.
The choice of contact threshold significantly influences profit. A suboptimal threshold can lead to substantial profit loss.
After reaching the peak, both curves exhibit decreasing returns, indicating that contacting customers beyond a certain threshold becomes less profitable.
The plot highlights the importance of selecting an appropriate contact threshold for maximizing profit. The Random Forest model demonstrates superior performance in this context.

Metric	Logistic Regression	Random Forest
Optimal Threshold	0.20%	0.32%
Percentage of Customers to Contact	22.22%	14.97%

Table 9: Model Performance Metrics

The optimal threshold for Logistic Regression (0.20%) is lower than that of Random Forest (0.32%). This suggests that Logistic Regression is more inclined to classify instances as positive compared to Random Forest for achieving optimal profit.The percentage of cus-

tomers to contact is significantly lower for Random Forest (14.97%) compared to Logistic Regression (22.22%). This implies that Random Forest is more selective in targeting customers, potentially leading to higher efficiency and lower costs.