


# Data Visualization take home assignment :

## Marketing campaign analysis

	Faculté		
		des <b>sciences économiques</b> et de <b>gestion</b>	
	Université de Strasbourg		

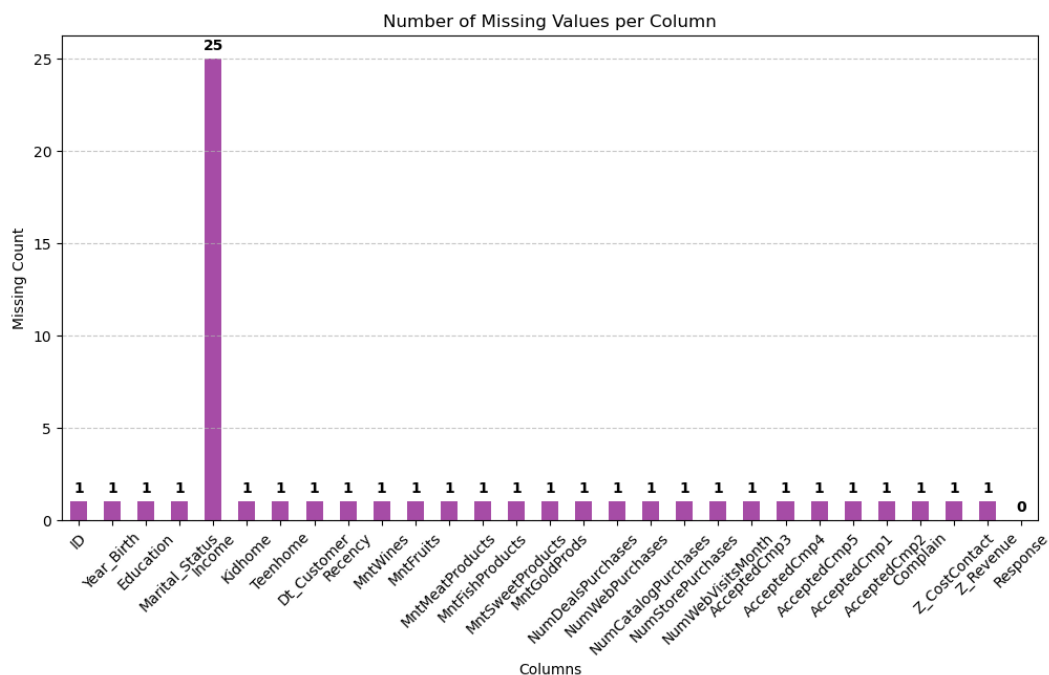
M2 DS2E 2024-2025  
FREZARD Paul  
SARTORI Romain  
VU Billy

## I) Data Insights

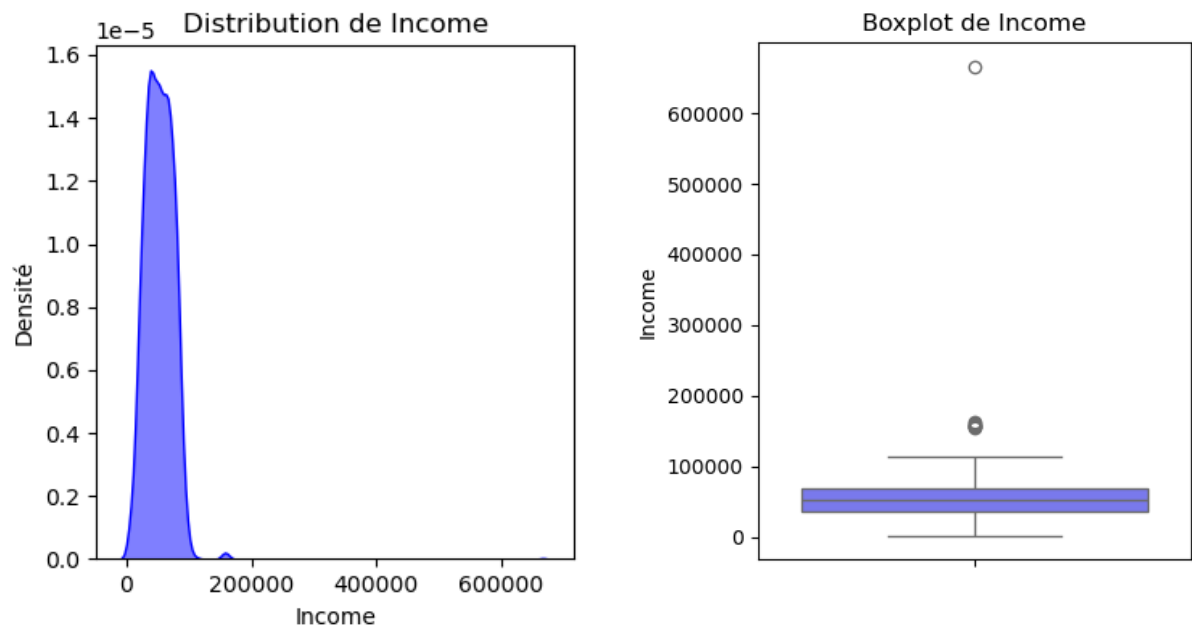
Before applying any predictive models, let's dive into the information and limitations of our original data set. To begin with, it would be interesting to get an overview of all the variables available.

The dataset includes customer demographics, purchasing behavior, and marketing interactions across 29 variables. Demographic data covers age, education, marital status, income, and household composition. Customer engagement is tracked through enrollment date and recency of purchases. Spending behavior is detailed across various product categories, while purchase channels include web, catalog, store, and promotional deals. Marketing interactions capture campaign responses, and customer satisfaction is reflected in complaint records and business metrics.

The dataset has minimal missing values, with most columns missing only **one value**, except for **Income**, which has **25 missing entries**. This suggests a **clean dataset** with minor gaps, requiring simple imputation strategies such as mean or median replacement for numerical fields.



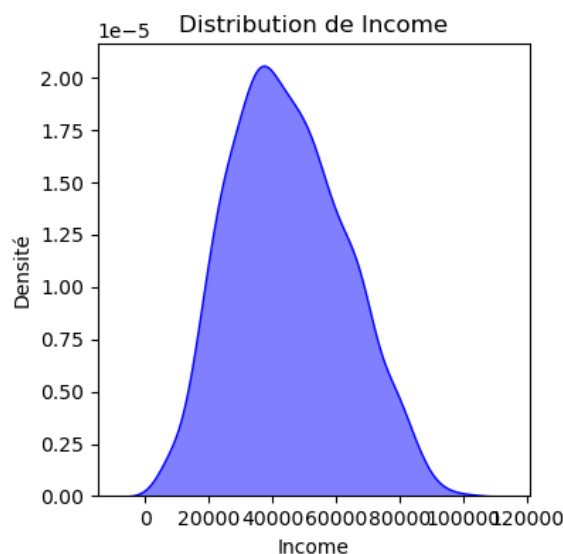
Now that we know what are the components of the data set. We wish to see into these components. We suggest two kind of plots : the a Kernel Density Estimate (KDE) plot, and the boxplot. Combining a KDE plot with a boxplot provides a more comprehensive view of a variable's distribution: while the KDE plot reveals the overall shape and density of the data, the boxplot highlights key summary statistics such as the median, quartiles, and potential outliers, making it easier to detect skewness, variability, and anomalies in the data.



The income distribution is highly unbalanced, with most people earning lower wages while a few earn significantly more. The density plot shows a sharp peak at lower incomes and a long stretch towards higher values, indicating a small group of high earners. The boxplot confirms this by highlighting several outliers and a wide gap between the majority and top earners. This suggests income disparity, which could impact pricing strategies, market segmentation, or salary structures.

Since we noticed the outliers presence on a regular basis throughout the study of all variables, we manage to start over with the above plots. But this time, we put aside the outliers to have a clearer sight of the distribution.

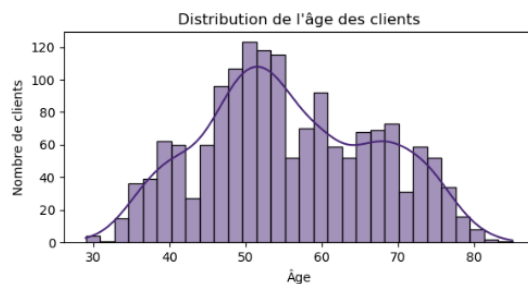
Again, we stick with our example : the income distribution.



With outliers removed, the income distribution is more balanced, forming a bell-shaped curve with most incomes clustered around the middle range. This suggests a more homogeneous market, where the majority of people have similar purchasing power. Unlike the previous skewed distribution, this refined view helps businesses better segment their audience, set realistic pricing strategies, and design products or services that cater to the core customer base rather than being influenced by extreme high earners.

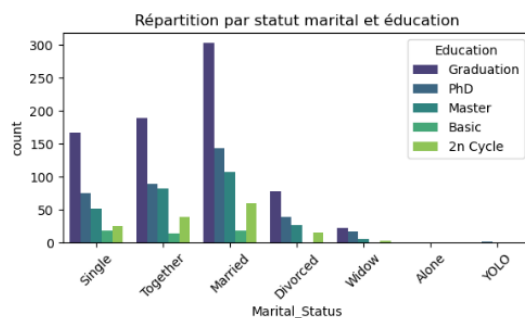
Let's have a look at others variables distribution :

### Age distribution :



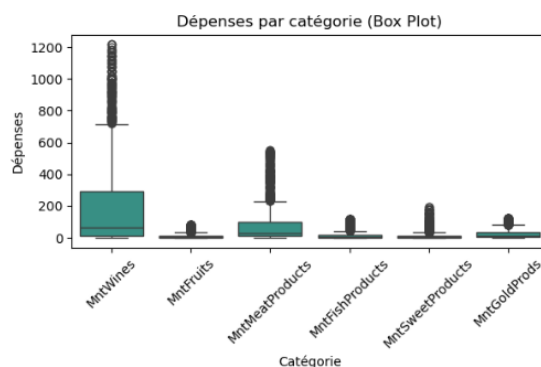
This histogram reveals the age distribution of customers, showing a balanced spread with a peak around 40-60 years old. This suggests that most customers belong to the middle-aged segment, which can influence marketing strategies, product offerings, and communication style.

### Marital Status & Education Breakdown



This bar chart provides insights into the demographics and education levels of customers. The majority appear to be married or in a relationship, suggesting that family-oriented marketing strategies could be effective. Education levels may influence buying behaviors, brand perception, and product preferences, particularly for premium or specialized products.

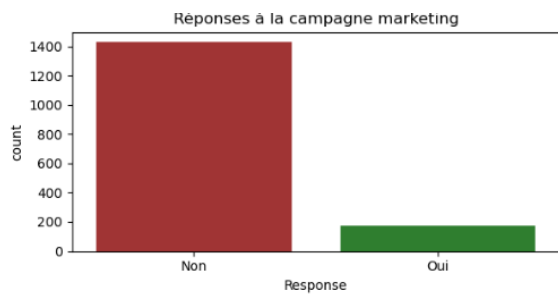
### Spending by categories



This box plot shows spending patterns across different product categories, identifying which products drive the most revenue. The wine category (MntWines) has the highest median spending and most outliers, indicating a strong interest in this product. Other categories show lower spending and less variability, suggesting potential areas for marketing improvement or product bundling strategies.

Identifying high-spending categories helps optimize inventory, pricing, and promotional campaigns.

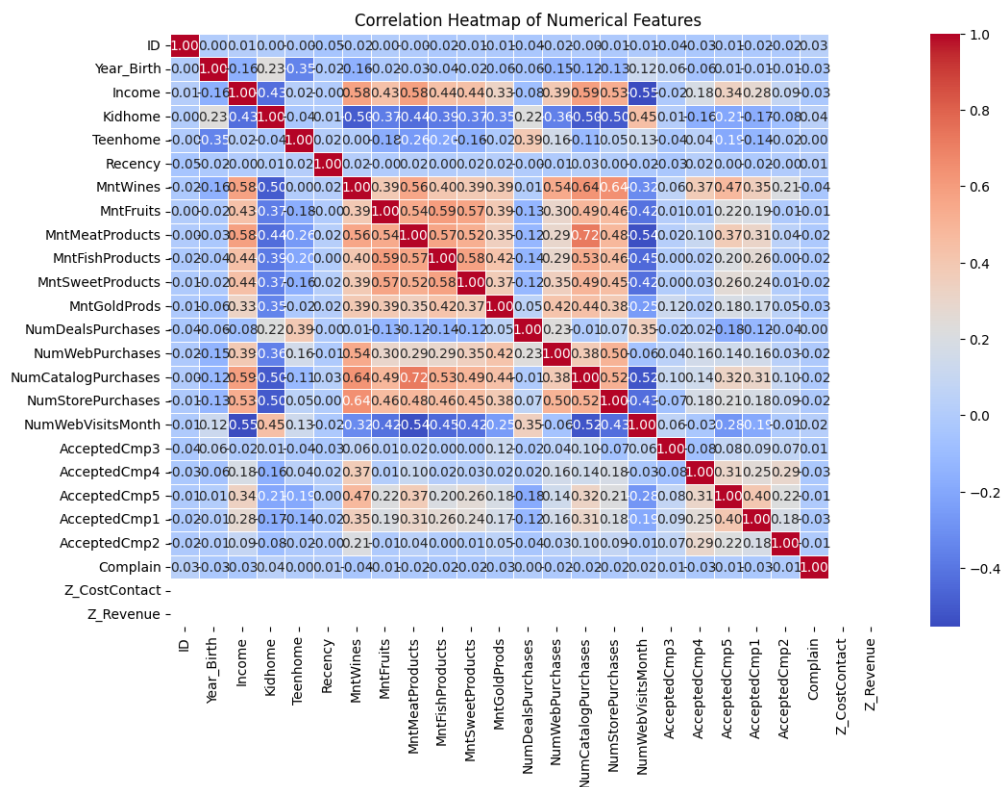
## Marketing Campaign Response



The marketing campaign response plot above indicates customer engagement with marketing campaigns, useful for evaluating campaign effectiveness. This bar chart shows that the majority of customers did not respond to the marketing campaign, with only a small fraction engaging. A low response rate suggests a need

for better-targeted marketing strategies, improved messaging, or alternative communication channels.

Now that we have studied several variables inside the data set, it would be interesting to see how they interact with one another.

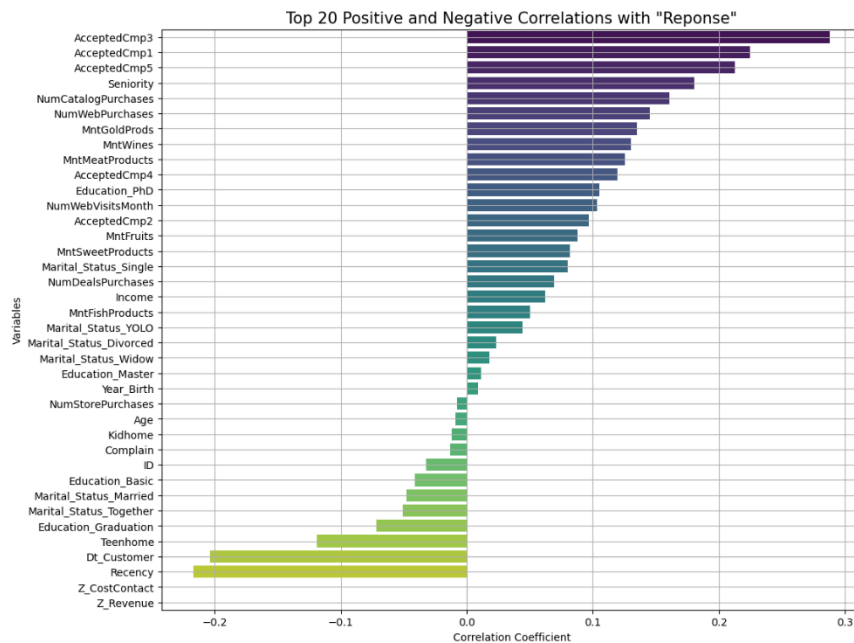


The heat map reveals key insights:

- Strong Negative Correlation: Year of birth and income (-0.16) indicate younger individuals tend to earn less.
- Strong Positive Correlation: Consumption categories correlate highly (above 0.50), meaning higher spenders buy across multiple categories. Web visits and purchases (0.50) suggest frequent visitors shop more online.

- Moderate Positive Correlation: Income moderately correlates (0.4–0.5) with total spending, showing wealthier individuals spend more.
- Weak or No Correlation: Complaints and past campaign acceptance show little impact on purchasing behavior.

We could also see more specifically the correlation of all the variables to the response one.



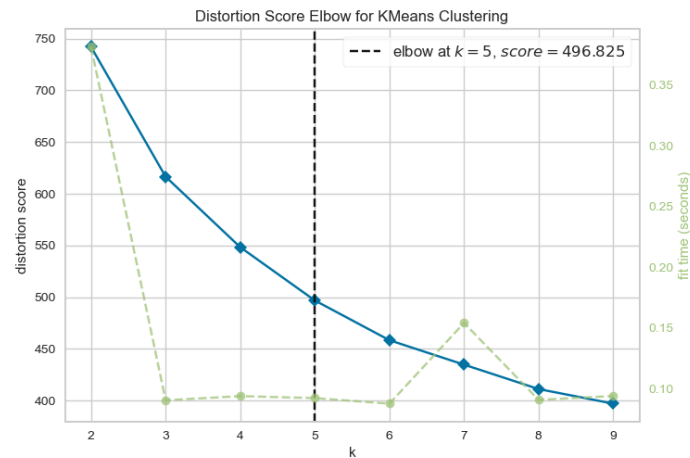
The bar chart highlights the top 20 positive and negative correlations between various factors and customer response. The strongest positive correlations are linked to previous campaign acceptances, indicating that customers who engaged in past campaigns are more likely to respond again. On the negative side, lower total spending, higher costs per contact, and more recent interactions show the strongest negative correlations, suggesting that customers who spend less overall, require higher marketing costs, or were recently contacted are less likely to respond. This visualization helps identify key drivers of customer engagement.

## II) Market Segmentation: Enhancing Targeted Marketing Strategies

In any marketing campaign, understanding customer differences is crucial for optimizing Market segmentation is a crucial step in understanding customer behavior and optimizing business strategies. By leveraging Self-Organizing Maps (SOM) and K-Means clustering, we have identified distinct customer segments that can inform targeted marketing actions. Below is an analysis of the segmentation process and key insights derived from it.

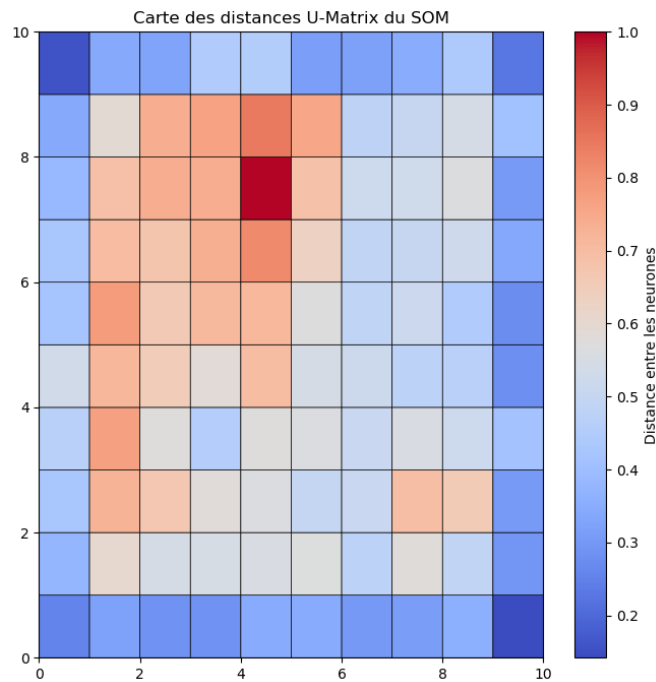
## Identifying the Optimal Number of Clusters

The Elbow Method (graph below) was used to determine the ideal number of customer segments. The analysis suggests that five clusters provide the best balance between interpretability and differentiation. This ensures that each segment is distinct enough to be actionable while avoiding excessive complexity.



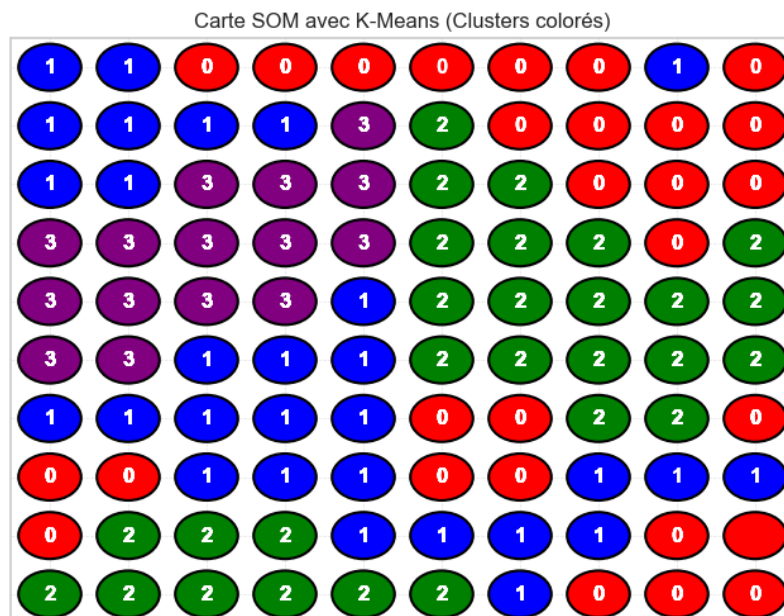
## SOM Distance Matrix – Understanding Customer Similarities

The U-Matrix of the SOM (graph below) visually represents the distances between customer groups. Darker colors indicate areas where customers are closely related, while lighter or red-colored zones show greater dissimilarity. This highlights natural divisions in the dataset, confirming that distinct customer groups exist and justifying the clustering approach.



### Cluster Visualization with K-Means and SOM

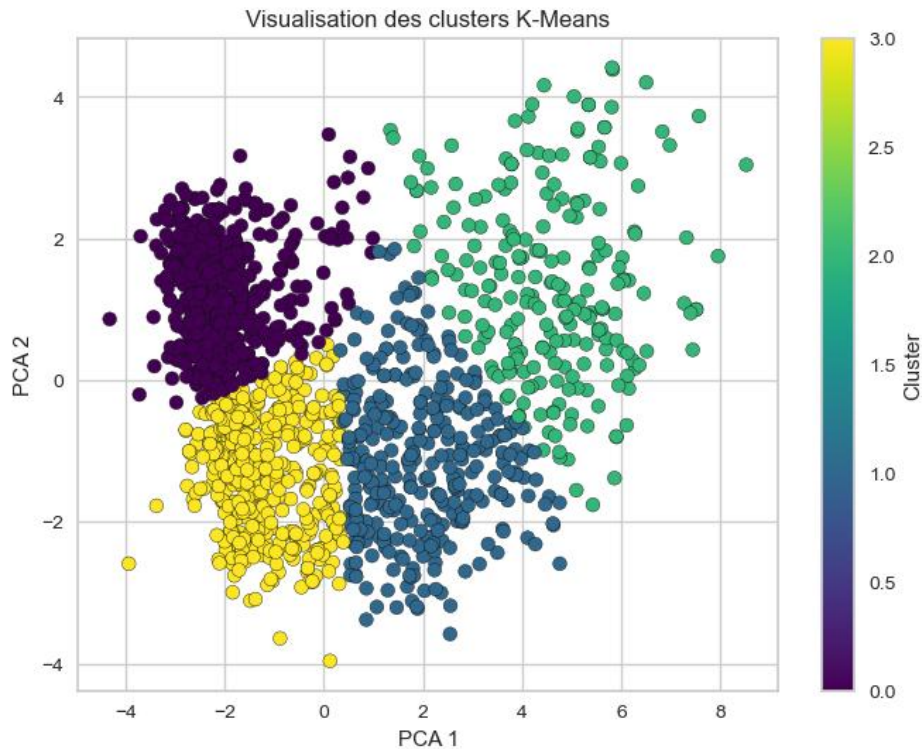
The SOM cluster visualization (second graph) provides an intuitive representation of how customers are mapped onto the SOM grid. Each number and color represents a different cluster, revealing how different segments are spread across the customer base.



Similarly, the PCA-based K-Means clustering visualization (fourth graph) projects the identified segments onto two principal components, making it easier to see how groups



separate in a lower-dimensional space. The distinct clusters suggest that customers exhibit varied behaviors, purchasing patterns, or engagement levels.



### Business Implications of Market Segmentation

This segmentation provides several actionable insights:

- High-value customers (e.g., premium spenders) can be identified and targeted with exclusive offers or loyalty programs.
- Price-sensitive segments may respond better to discount-based promotions.
- Online vs. in-store shoppers can be approached with channel-specific marketing campaigns.
- At-risk customers (low engagement) can be re-engaged through personalized communications or incentives.

## III) Comparison of Predictive Models: Logistic Regression vs. Random Forest

When evaluating predictive models for business applications, the key considerations include accuracy, precision, recall, and overall ability to differentiate between potential

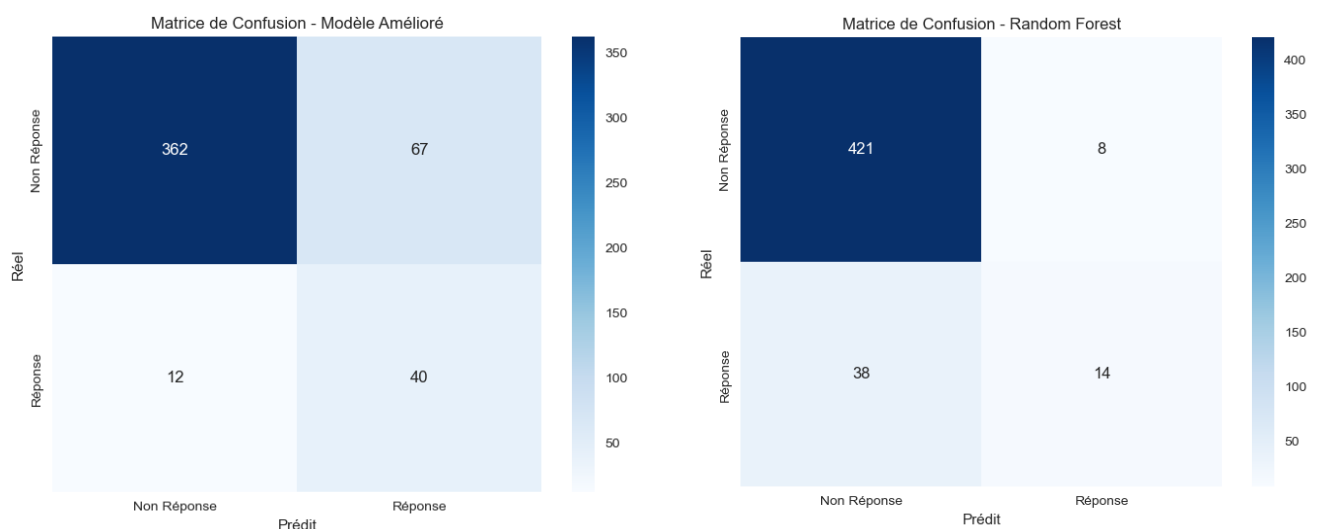
responders and non-responders. In this analysis, we compare Logistic Regression and Random Forest to determine which model is better suited for targeting customers.

Both models were evaluated using confusion matrices, ROC curves, and classification reports, which provide insights into their effectiveness.

### Confusion Matrix Analysis

- Logistic Regression correctly predicted 362 non-responders but misclassified 67 customers who did not respond. It captured 40 true responses while missing 12 potential responders.
- Random Forest showed a stronger distinction in classification, with 421 correct non-responders and only 8 misclassified. However, it only captured 14 true responses, misclassifying 38 responders as non-responders.

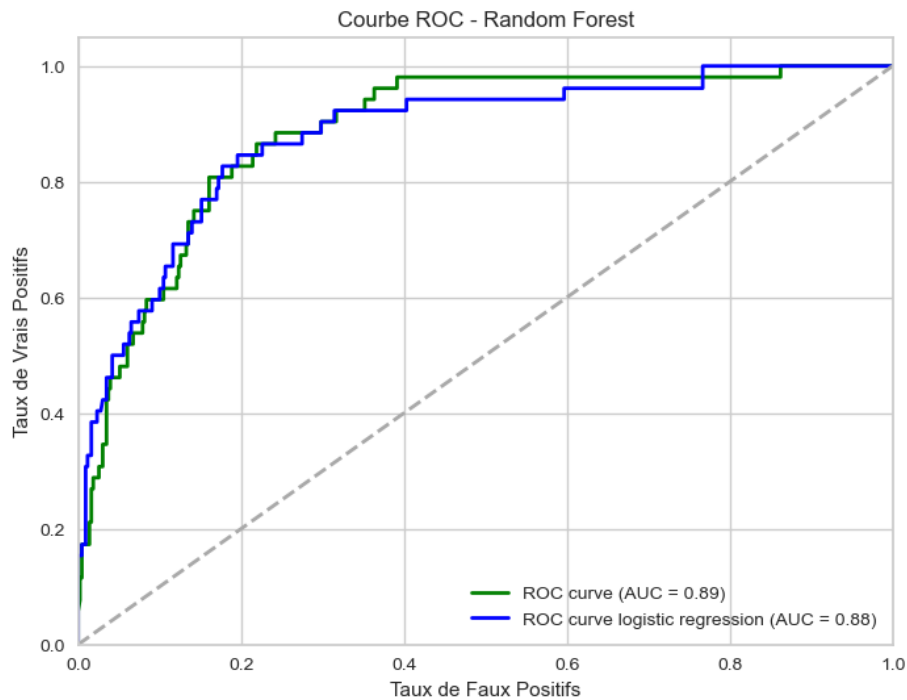
The Logistic Regression model does a better job at capturing potential responders (higher recall for the positive class), which is crucial for marketing campaigns where missing a potential conversion can be costly. The Random Forest model is more conservative, avoiding false positives but at the cost of missing many real responders.



### ROC Curve & AUC Comparison

- The Logistic Regression model achieved an AUC of 0.88, indicating strong predictive power.
- The Random Forest model slightly outperformed with an AUC of 0.89, showing a marginally better ability to differentiate responders from non-responders.

A higher AUC indicates a better-performing model overall, meaning Random Forest can better separate positive and negative cases. However, the difference is minimal, meaning both models are viable choices depending on business priorities.



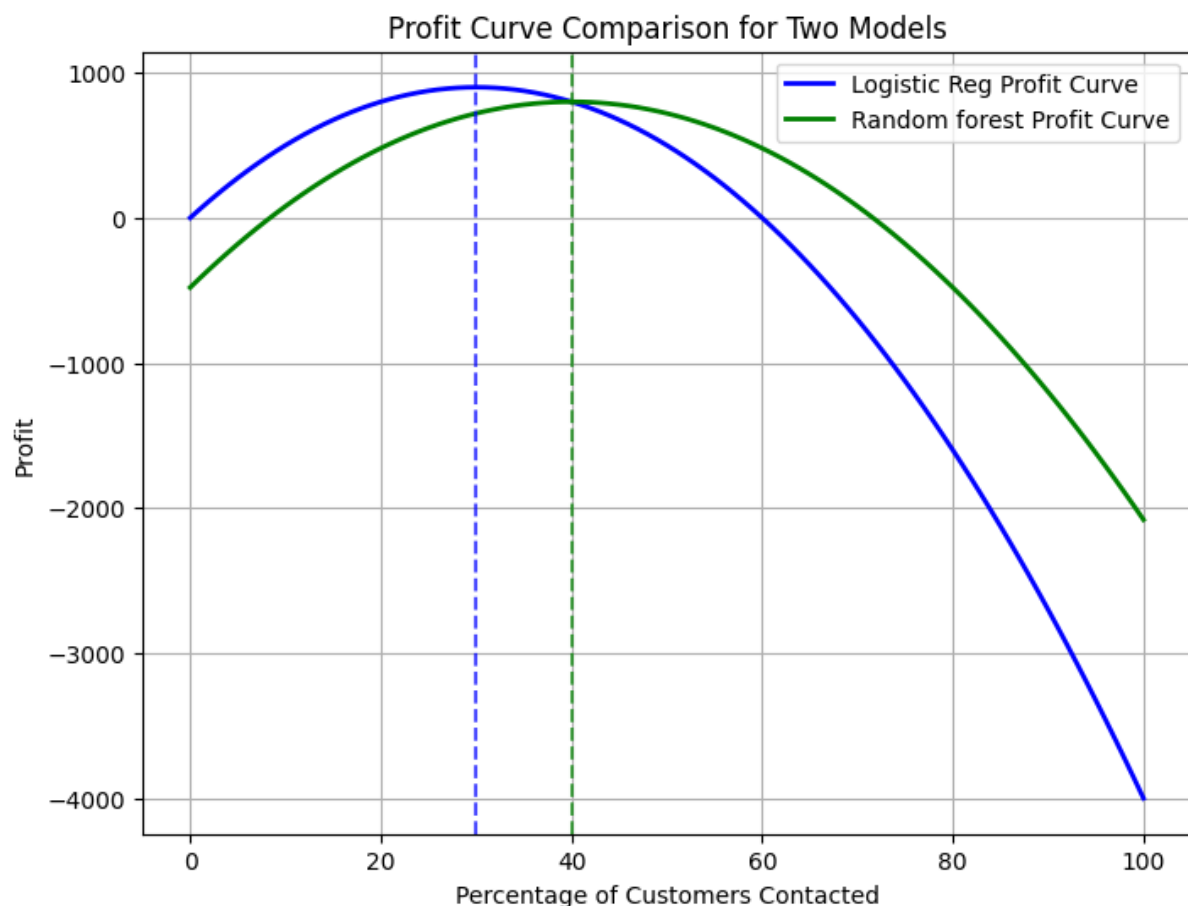
### Classification Report – Precision, Recall, and F1-Score

- Logistic Regression showed:
  - Higher recall for responders (0.27 vs. 0.77 for Random Forest), meaning it captures more actual responders.
  - A lower precision (0.64 vs. 0.37 for Random Forest), meaning it has more false positives.
  - Lower F1-score (0.38 vs. 0.50 for Random Forest), showing that, overall, Random Forest has a better balance between precision and recall.
- Random Forest:
  - High recall for non-responders (0.97) but low recall for responders (0.37), meaning it struggles to detect actual conversions.
  - Higher F1-score (0.50 vs. 0.38 for Logistic Regression), suggesting a better trade-off between recall and precision.

If the goal is to maximize customer responses, Logistic Regression is the better choice, as it identifies more potential responders, even at the cost of some false positives. If the goal is to avoid wasted marketing efforts on unlikely responders, Random Forest is better, as it reduces false positives.

## IV) Comparison of the Models Based on the Profit Curve

The profit curve :



The profit curve analysis reveals the optimal percentage of customers to contact to maximize profitability. Both models, Logistic Regression and Random Forest, show an initial increase in profit as more customers are contacted. However, after a certain threshold, profits begin to decline as the cost of outreach surpasses the revenue generated.

Logistic Regression reaches its maximum profit at around 28% of customers contacted, making it a suitable choice for a cost-efficient marketing strategy where minimizing expenses is a priority. In contrast, Random Forest achieves its highest profit at 42%, suggesting a broader customer outreach is viable while still remaining profitable. Although the absolute profit is slightly higher for Random Forest, it requires a larger budget to reach its optimal performance.

From a business perspective, the decision depends on the company's marketing strategy and budget constraints. If the goal is to maximize returns while keeping costs low, targeting 28% of customers using Logistic Regression is the most effective approach. However, if the company has the financial capacity for a more extensive campaign, reaching 42% of customers with Random Forest could yield a higher overall profit.

A hybrid approach may also be beneficial, where the company initially targets the most promising customers using Logistic Regression and later expands the outreach using Random Forest's threshold if budget permits. This ensures a balanced strategy, maximizing profit while controlling costs and optimizing resource allocation.