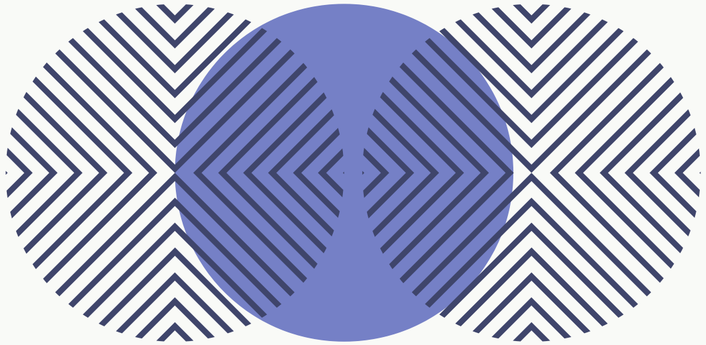
BUSINESS PRESENTATION

OPTIMIZING THE PROFITABILITY OF MARKETING CAMPAIGNS USING DATA ANALYSIS

How can we identify high-potential customers to maximize the success of direct marketing campaigns?



Oumou SOW
Lath ESSOH
Tsiba RAZAFINDRAKOTO

M2 DS2E
Unistra

CONTENTS

01

INTRODUCTION & BUSINESS CONTEXT

Market Overview

BACKGROUND & BUSINESS CHALLENGE

Marketing campaigns are essential for businesses to attract and retain customers. However, many companies struggle with **low response rates and high acquisition costs**, leading to poor profitability. Traditional mass-marketing strategies often fail to reach the right audience, resulting in wasted budget on customers who are unlikely to convert.

A previous marketing campaign, aimed to sell a new gadget to customers and targeting **2,240 customers** had a **success rate of only 15%**, generating a revenue of **3.674MU** against a cost of **6.720MU**, leading to a total **loss of -3.046MU**. To address this issue, the objective of this project is to develop a **data-driven approach** that helps businesses:

- Identify high-potential customers most likely to respond to a marketing offer.
- Optimize targeting to reduce wasted marketing costs.
- Maximize campaign profitability through predictive modeling.

PROJECT OBJECTIVES & KEY QUESTIONS

Our goal is to **build a predictive model that will increase the profitability of future marketing campaigns** by analyzing past customer data and purchasing behavior. Specifically, we aim to answer:

- How can customer segmentation improve marketing efficiency?
- What are the key features that influence customer purchasing decisions?
- Which predictive models can best estimate the likelihood of a customer responding to a campaign?
- What is the optimal percentage of customers to target to maximize profits?

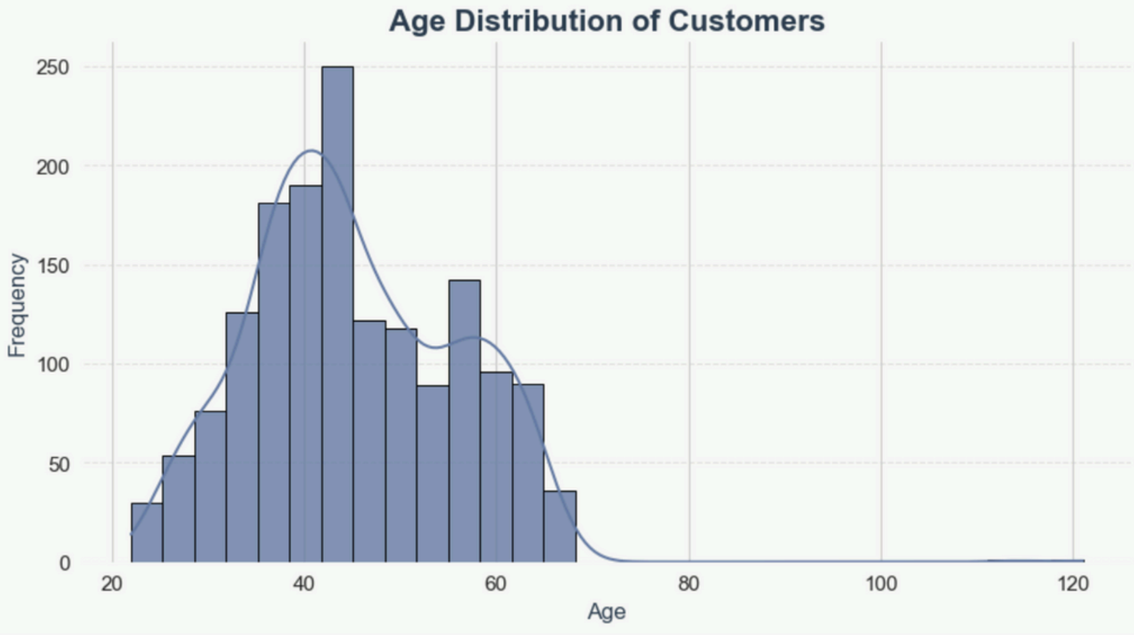
02

EXPLORATORY DATA ANALYSIS

1- Customer Demographics: Who are the customers?

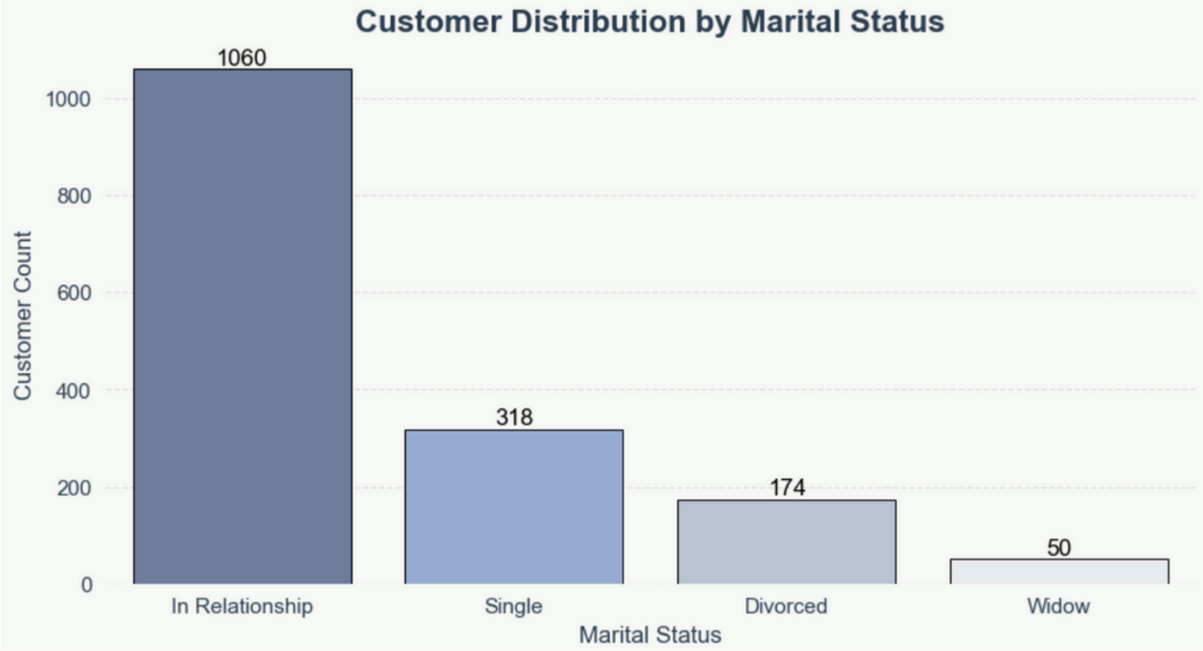
AGE DISTRIBUTION

- The majority of customers are aged between 30 and 60 years old.
- A small fraction of customers are over 70, indicating potential for a senior-focused campaign.



MARITAL STATUS DISTRIBUTION

- Most customers are in a relationship.
- Single customers are the second-largest group.
- Few customers are divorced or widowed, possibly influencing spending habits.



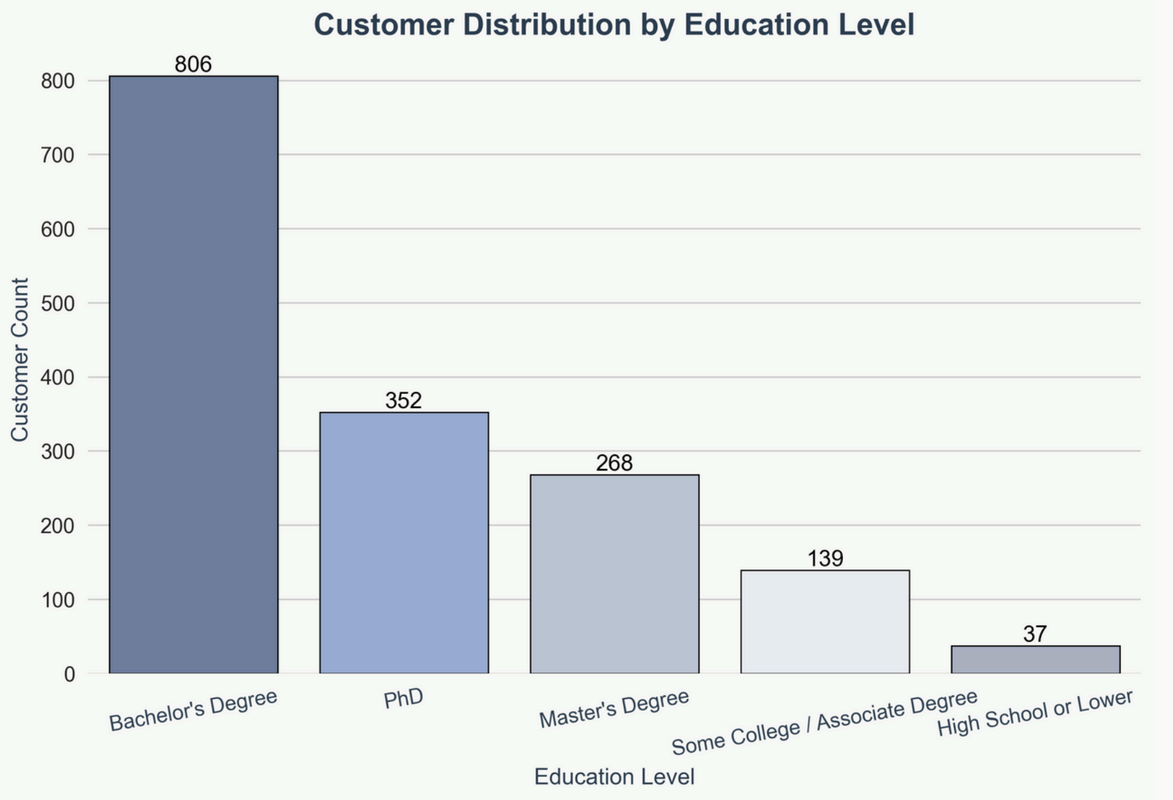
02

EXPLORATORY DATA ANALYSIS

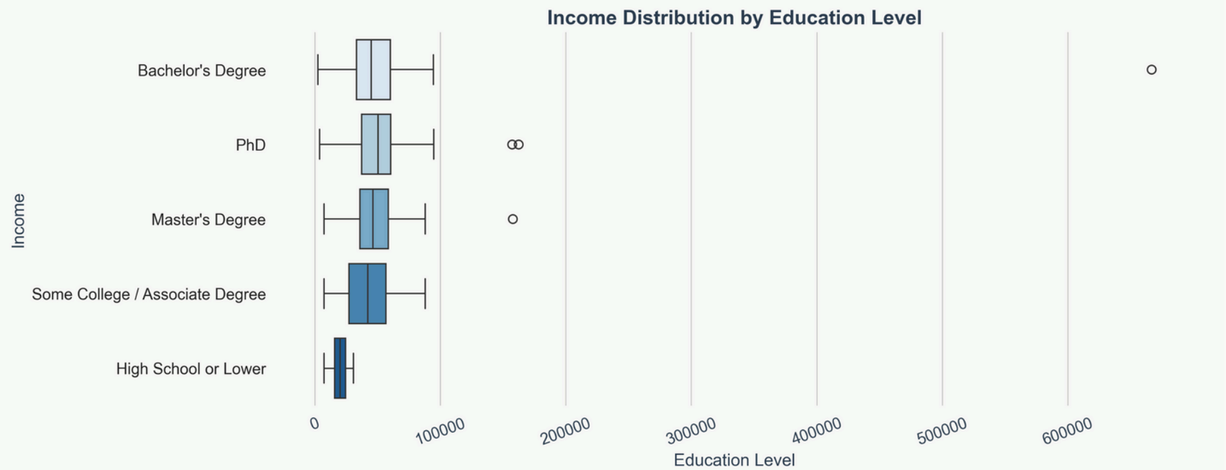
1- Customer Demographics: Who are the customers?

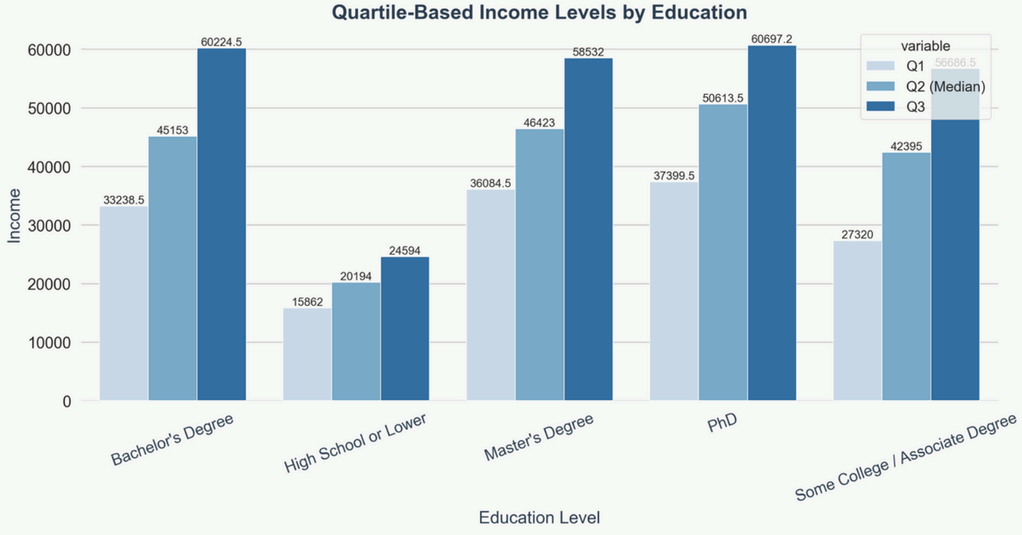
EDUCATION LEVEL

- A large proportion of customers hold a Bachelor's Degree.
- Customers with PhDs and Master's degrees form a significant portion.
- Lower education levels correspond to fewer customers, which may correlate with income levels.



INCOME DISTRIBUTION BY EDUCATION LEVEL





02

EXPLORATORY DATA ANALYSIS

1- Customer Demographics: Who are the customers?

BUSINESS INSIGHTS FROM INCOME AND EDUCATION ANALYSIS

Income Quartiles by Education Level

This chart presents income distribution by education level across three quartiles:

- Q1 (25th percentile): The lower range of typical income in this group.
- Q2 (Median, 50th percentile): The middle income value for each education level.
- Q3 (75th percentile): The upper range, representing higher earners in each category.

Observations:

- Individuals with a PhD or Master's degree earn significantly more on average compared to those with lower educational levels.
- Bachelor’s degree holders show a relatively high median income, making them an interesting target segment.
- High school graduates and those with some college education have the lowest income, with limited dispersion, indicating financial constraints in consuming premium products.

Income Distribution by Education Level

This boxplot provides a more detailed view of income variation and highlights outliers across education levels.

Observations:

- High variability in PhD and Master’s Degree holders' income: These groups exhibit a wide income distribution, meaning they include both high earners and those in lower-income brackets.
- High School Graduates' income is more concentrated, with low variation, suggesting wage stagnation within this group.
- Outliers among PhD and Master’s Degree holders: Some individuals have exceptionally high incomes, possibly representing entrepreneurs, executives, or industry experts.

Business Implications:

- Higher Education = Greater Financial Stability: they have higher purchasing power and are less impacted by economic downturns. Instead of mass-marketing, prioritizing highly educated customers (Master’s, PhDs, Bachelor's) could improve the campaign's efficiency.
- Middle-Educated Consumers Show Income Variance: Income levels vary within this group, requiring segmentation beyond education. A one-size-fits-all strategy will not work.
- A data-driven segmentation strategy can optimize marketing spend, ensuring budget is directed toward the highest-converting customers. So the new approach is to use predictive modeling to focus on high-probability buyers.

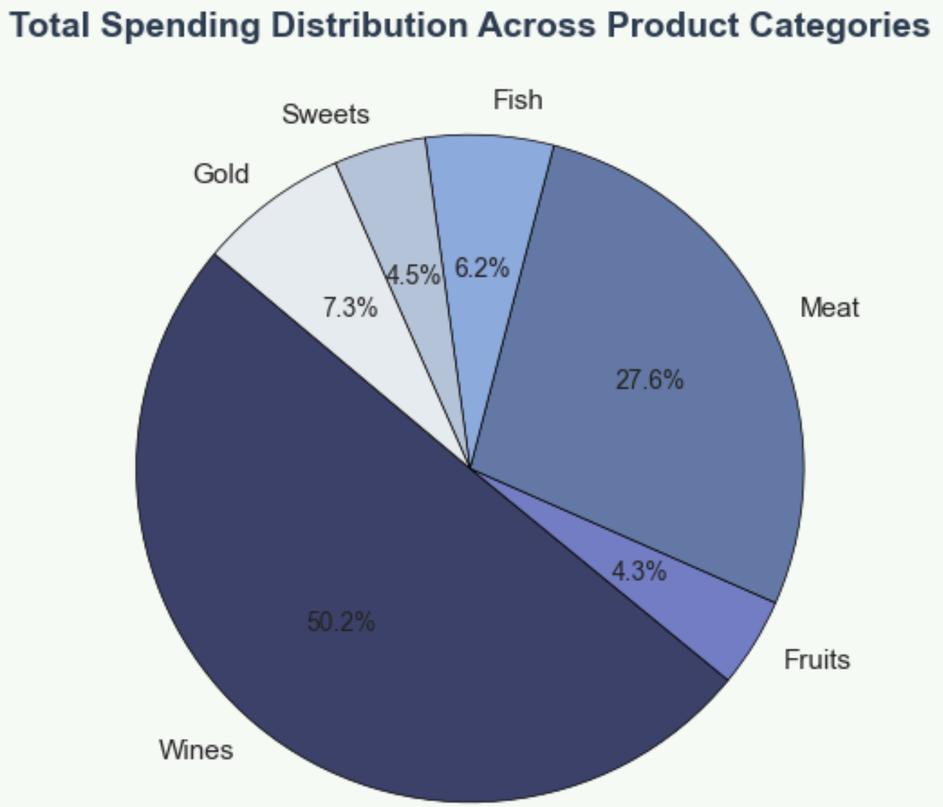
02

EXPLORATORY DATA ANALYSIS

2- Spending Behavior: What do customers buy? Which categories are most popular?

Observations:

- Wines dominate customer spending, accounting for nearly half (50.2%) of total purchases. This suggests that wine-related promotions, loyalty programs, or premium wine selections could be a major revenue driver.
- Meat products are the second-largest category (27.6%), indicating strong demand for high-quality food items. Bundling offers with wines or other gourmet products may enhance sales.
- Gold products (7.3%) represent a niche but significant luxury segment, suggesting that some customers have high disposable income and could be targeted for premium or high-end product campaigns.
- Fruits (4.3%), sweets (4.5%), and fish (6.2%) have lower spending shares, but they could be complementary products in personalized marketing strategies (e.g., promoting seafood and wine together).



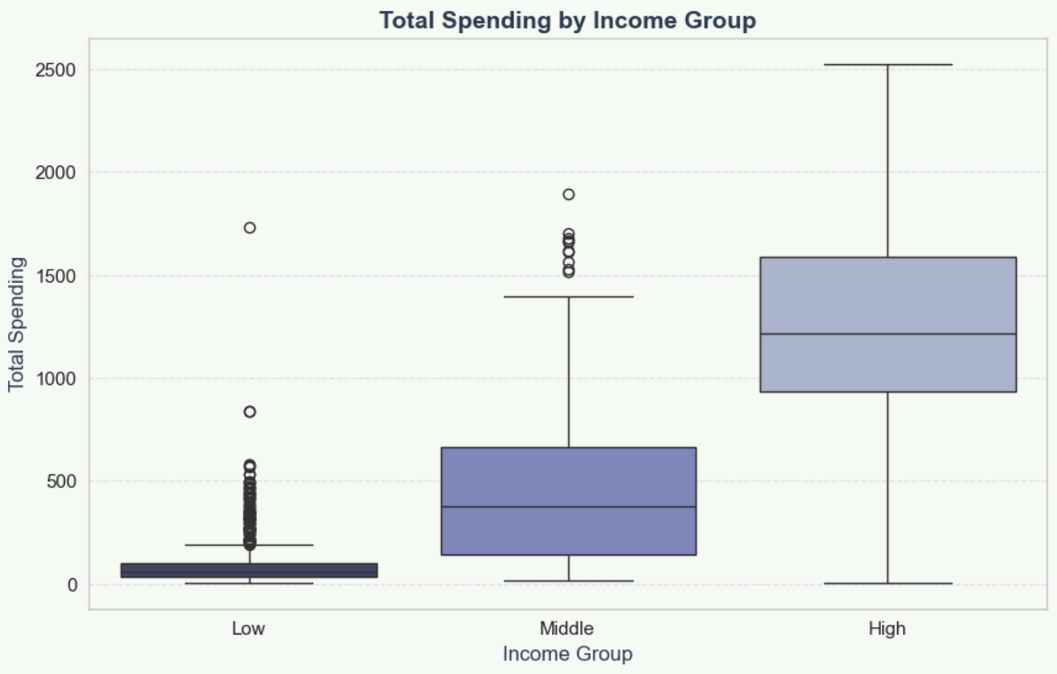
Business implications:

- Target premium customers with wine and gourmet product promotions, as these categories drive the most spending.
- Segment marketing strategies: Offer exclusive discounts for high-end spenders (gold products) and targeted campaigns for mass-market products like meat and seafood.

02

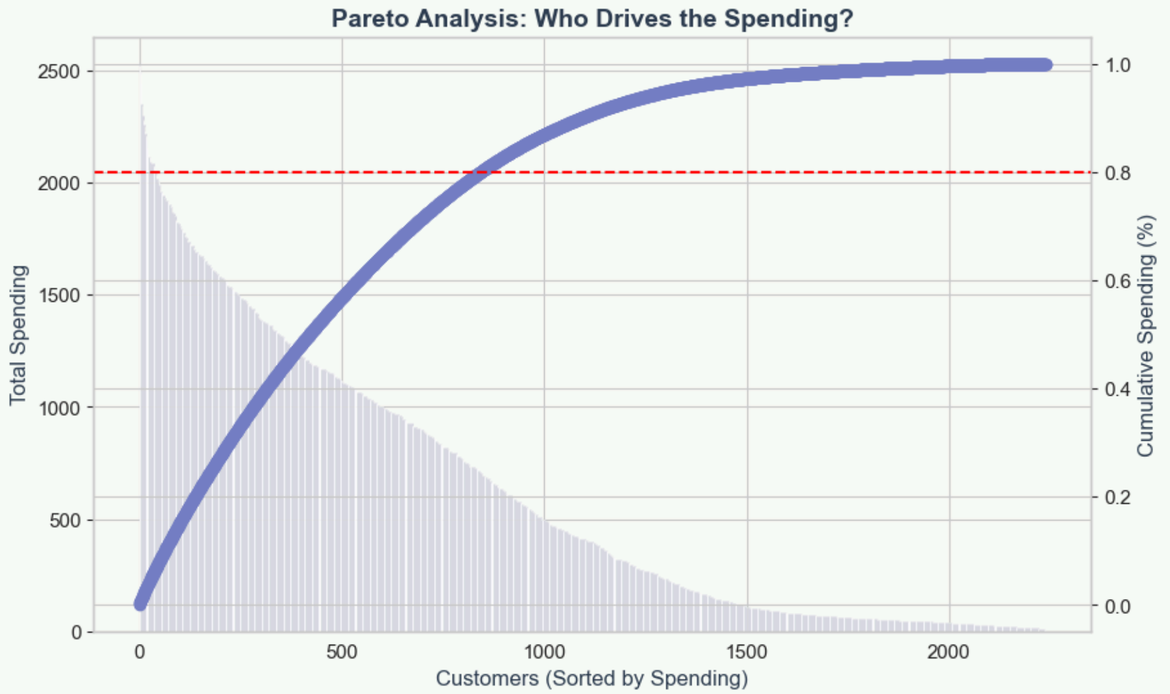
EXPLORATORY DATA ANALYSIS

3- Income vs. Spending: Do high-income customers spend more? What's the best target segment?



Observations:

- The boxplot divides customers into three income groups: Low, Middle, and High.
- Higher-income customers tend to spend significantly more than low and middle-income customers
- There is a wider distribution of spending in high-income customers, meaning some individuals spend much more than the average, while others spend conservatively.
- The low-income group has minimal spending, with very few outliers showing higher spending.
- Middle-income customers display a mix, with moderate spending habits but a few individuals who spend considerably.



- The Pareto principle (80/20 rule) is evident, where approximately 20% of customers contribute to 80% of the total spending.
- The top spenders contribute disproportionately to revenue, while a long tail of lower spenders generates much smaller contributions.
- The cumulative spending curve (blue) reaches 80% around the top 20% of customers, showing that targeting a smaller high-spending group is more effective than a broad marketing strategy.

High-income customers have significantly higher spending, but spending behavior is not uniform within each income group.

02

EXPLORATORY DATA ANALYSIS

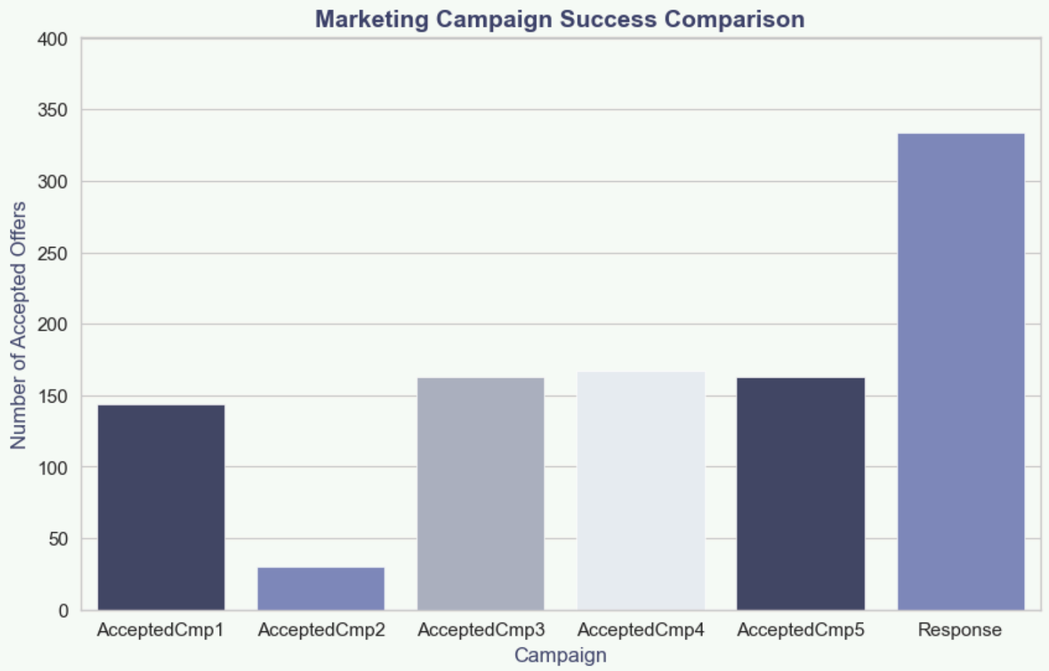
4- Purchasing Channels: Where do customers shop? (Online, Store, Catalog, Deals)

- The highest number of purchases occur in physical stores, indicating that a large segment of customers prefer in-person shopping.
- Online purchases (Web) rank second in total purchases, showing that a significant portion of customers is adopting e-commerce.
- Catalog purchases are less frequent, indicating that traditional mail-order shopping may be declining.
- The least number of purchases come from discount-based deals

02

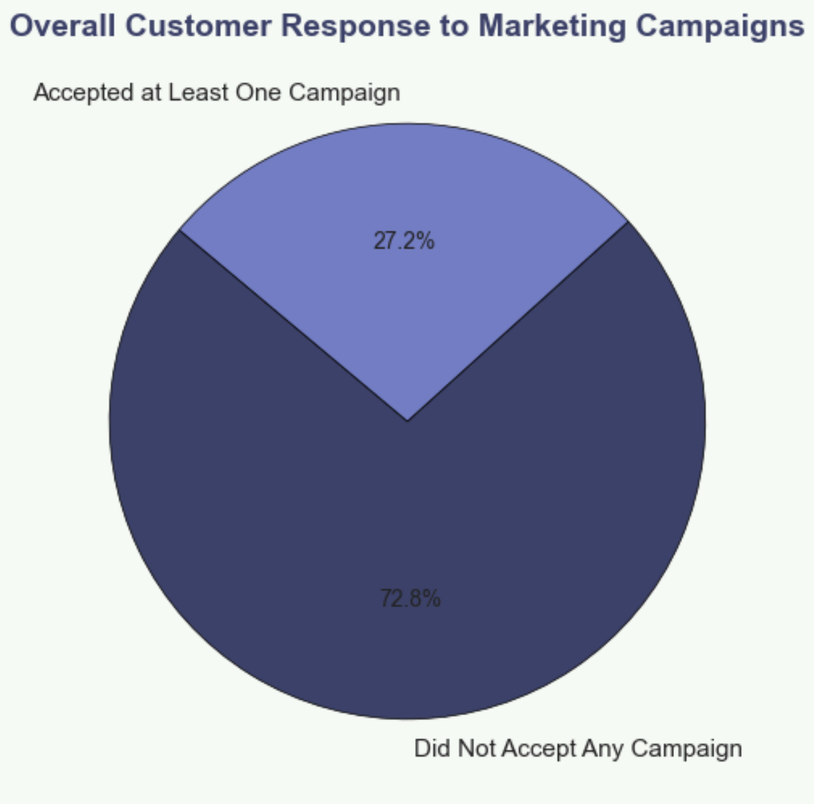
EXPLORATORY DATA ANALYSIS

5- Marketing Campaign Success: Which campaigns were effective? How to improve future campaigns?



- Only 27.2% of customers accepted at least one campaign, while 72.8% rejected all offers.
- This suggests that customer engagement with marketing efforts is relatively low.

There may be a need for better segmentation, personalized offers, or improved communication strategies.



Observations:

- Campaign 2 had the lowest acceptance rate, indicating it was the least effective.
- Campaign 5 and the last response campaign had the highest acceptance rates, suggesting that either better targeting or improved campaign strategies were used.
- Campaigns 1, 3, and 4 had moderate success, with similar levels of accepted offers.
- The final campaign ("Response") had the most accepted offers, which could imply that insights from previous campaigns helped refine targeting strategies.

Future campaigns should analyze what worked in Campaign 5 and the final campaign to replicate their success.

03

MARKET SEGMENTATION USING SELF-ORGANIZING MAPS (SOM)

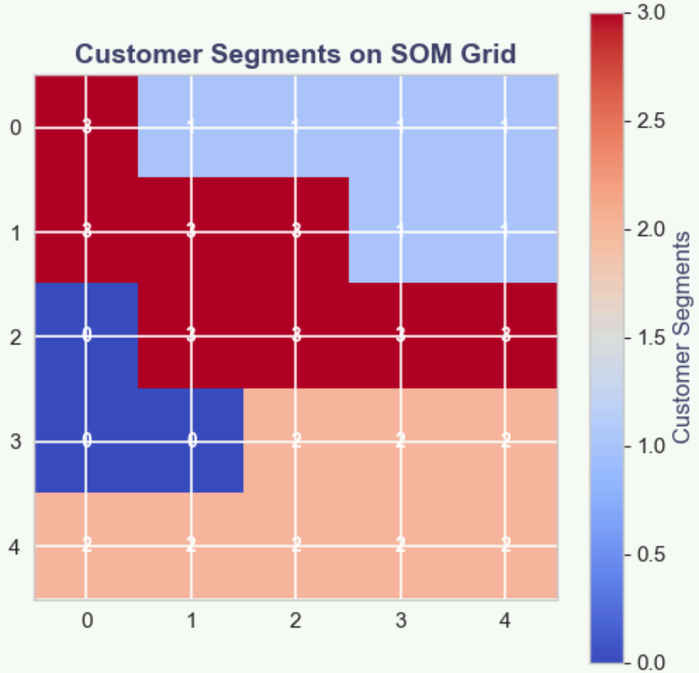
1- Understanding Customer Segments

Why Market Segmentation?

- To optimize marketing efforts and personalize customer interactions, we identified key customer groups.
- Grouping similar customers helps in better targeting and resource allocation.

How We Did It?

- We used Self-Organizing Maps (SOM), an advanced AI clustering technique, to analyze purchasing behavior and customer demographics.
- Unlike traditional segmentation, SOM detects hidden patterns automatically.



03

MARKET SEGMENTATION USING SELF-ORGANIZING MAPS (SOM)

2- Customer Segments Breakdown: Who Are They?

Segment 0: Store-Oriented Shoppers

- Profile: High-income customers who prefer in-store purchases.
- Behavior: They make frequent purchases, especially from physical stores.
- Strategy: Encourage store loyalty with exclusive in-store offers and premium services.

Segment 1: Budget-Conscious Customers (Largest Group)

- Profile: Low-income customers with minimal spending.
- Behavior: Rarely purchase online or in-store.
- Strategy: Discount promotions, free trials, and referral incentives to increase engagement.

Customer Segment Summary (mean)

	Income	NumWebPurchases	NumStorePurchases	Total_Spending
0	66252.13	8.09	9.09	1128.99
1	32479.06	1.95	3.02	78.46
2	77912.65	4.86	8.24	1422.99
3	54466.15	5.18	6.62	543.91

Segment 2: Selective Premium Buyers

- Profile: Highest-income customers, selectively spending on high-quality products.
- Behavior: Engage with online and store purchases but only for premium items.
- Strategy: Target with VIP programs, premium memberships, and high-end product promotions.

Segment 3: Digital-First Customers

- Profile: Mid-income customers, mainly shopping online.
- Behavior: Prefer web purchases but also the physical stores
- Strategy: Strengthen online experience with personalized recommendations, flash sales, and digital loyalty programs.

04

PREDICTIVE MODELING

For prediction, we have chosen two high-performance models such as Random Forest Classifier for its robustness to noisy data and non-linear relationships and the fact that it handles categorical and numerical variables well without requiring scaling. The second model we will use is Gradient Boosting (XGBoost) for its excellent performance on classification problems and its ability to manage class imbalances and optimise errors iteratively.

1

Feature Selection

Select relevant variables to explain customer response.

- x : Selected features.
- y: Target variable (Response)

2

Data Splitting

- Use train_test_split to split data into 70% training and 30% test sets.
- The stratify=y parameter ensures a balanced distribution of classes.

3

Class Distribution Analysis

We observe a significant imbalance between the two classes. Class 0 is predominant compared to Class 1, which must be considered to avoid biasing the prediction.

PREDICTIVE MODELING

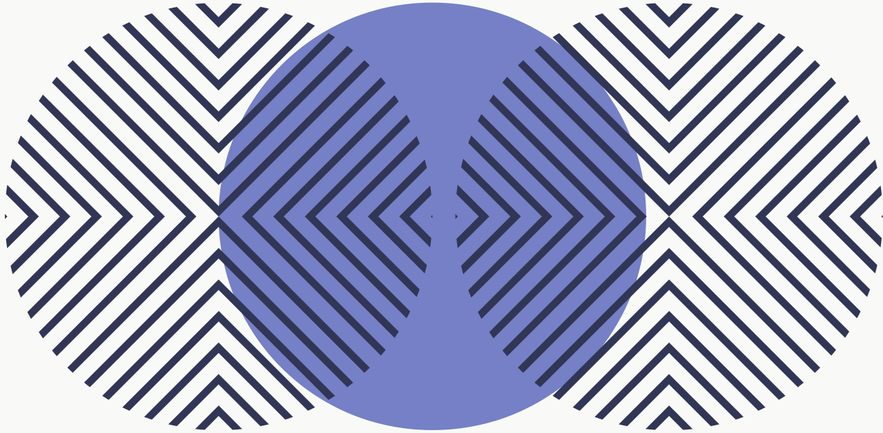
Given the class imbalance, we applied several techniques to improve the model’s performance

Hyperparameter Optimization

- Used GridSearchCV to test different hyperparameter combinations on a RandomForestClassifier
- Included class_weight='balanced' to better handle class imbalance.

Threshold Adjustment

Instead of the default 0.5 threshold, we adjusted it to 0.3 to improve the detection of the minority class.

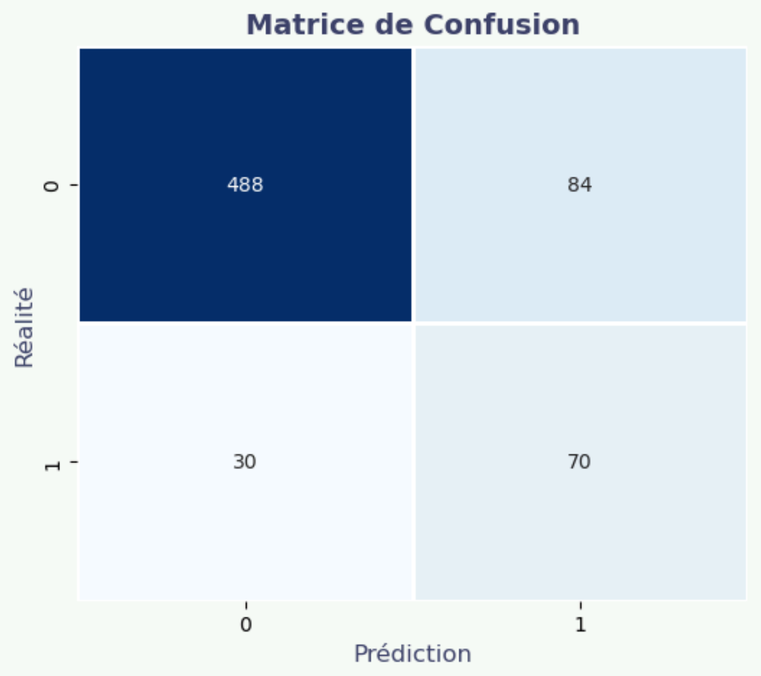


PREDICTIVE MODELING

Random Forest Classifier

Classification Report (Seuil Ajusté)

	precision	recall	f1-score	support
0	0.94	0.85	0.9	572.0
1	0.45	0.7	0.55	100.0
macro avg	0.7	0.78	0.72	672.0
weighted avg	0.87	0.83	0.84	672.0
accuracy			0.83	672.0



Performance Evaluation

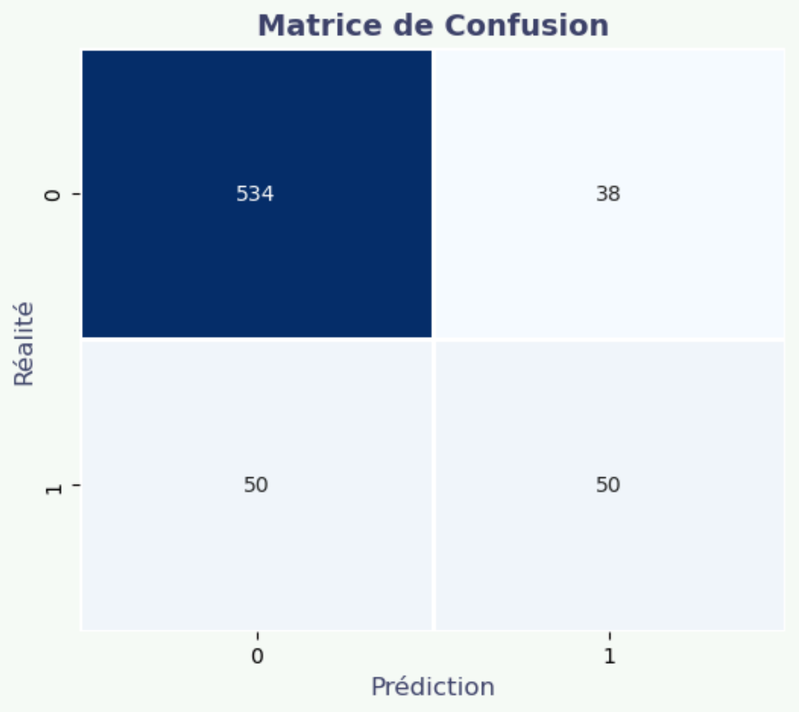
- The model has high precision (0.94) for non-buyers, effectively avoiding unnecessary costs by correctly identifying those not interested. However, the lower precision (0.45) for buyers leads to misdirected marketing efforts. With a recall of 0.70 for buyers, the model identifies a good portion of potential customers but still misses 30%, reducing sales opportunities. The significant gap in F1-scores (0.90 vs. 0.55) indicates that the model is much better at excluding non-buyers than detecting buyers. Finally, while the overall accuracy is 83%, this does not guarantee maximum profitability; improving buyer recall is crucial to optimizing the next campaign’s profits.

- The confusion matrix reveals that 488 true negatives help reduce communication costs by correctly identifying uninterested customers. However, 84 false positives increase marketing expenses without generating conversions. On the positive side, 70 true positives represent a secured profit opportunity, while 30 false negatives indicate a revenue loss, as these potential buyers were mistakenly classified as non-buyers and thus not targeted.

Gradient Boosting Classifier

Classification Report (Seuil Ajusté)

	precision	recall	f1-score	support
0	0.91	0.93	0.92	572.0
1	0.57	0.5	0.53	100.0
macro avg	0.74	0.72	0.73	672.0
weighted avg	0.86	0.87	0.87	672.0
accuracy			0.87	672.0



Performance Evaluation

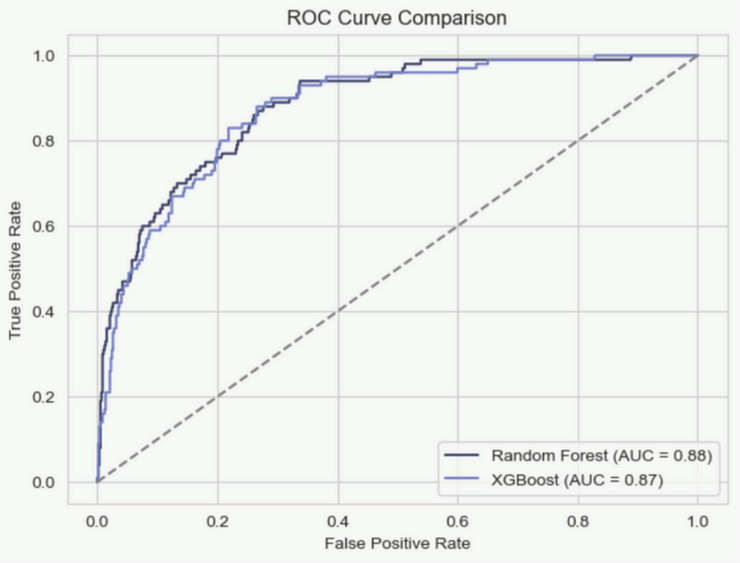
- The model has high precision for identifying non-buyers (0.91), helping to reduce unnecessary communication costs. However, with a precision of 0.57 for buyers, a significant portion of the predicted buyers are not actual buyers, leading to inefficient marketing expenses. The recall is high for non-buyers (0.93), but only 50% of actual buyers are correctly identified, posing a risk of lost sales opportunities. The gap between F1-scores (0.92 vs. 0.53) indicates that the model is much better at avoiding unnecessary expenses than maximizing sales. Despite an overall accuracy of 87%, the company must improve buyer detection to optimize the profitability of the marketing campaign.

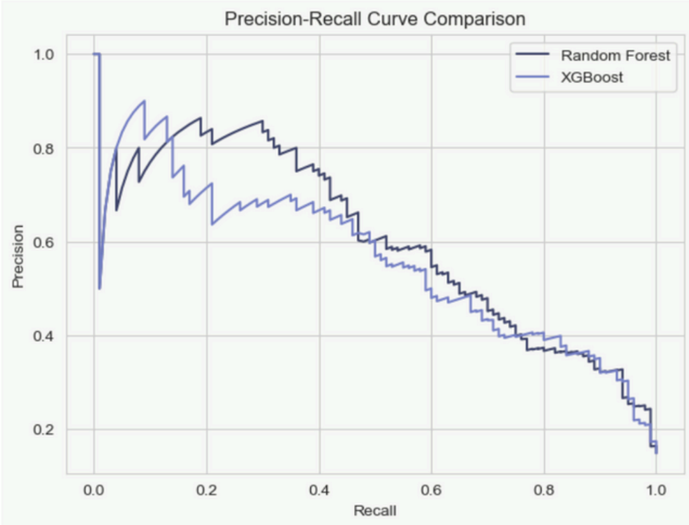
- The confusion matrix shows that 534 non-buyers were correctly identified, reducing costs associated with ineffective prospecting. However, 38 customers were mistakenly classified as buyers, leading to unnecessary marketing expenses without conversion. On the buyer side, 50 were correctly predicted, ensuring revenue through targeted prospecting. However, 50 potential buyers were misclassified as non-buyers, representing a loss of revenue since these customers could have been converted into sales.

04

PREDICTIVE MODELING

GRAPHICAL COMPARISONS





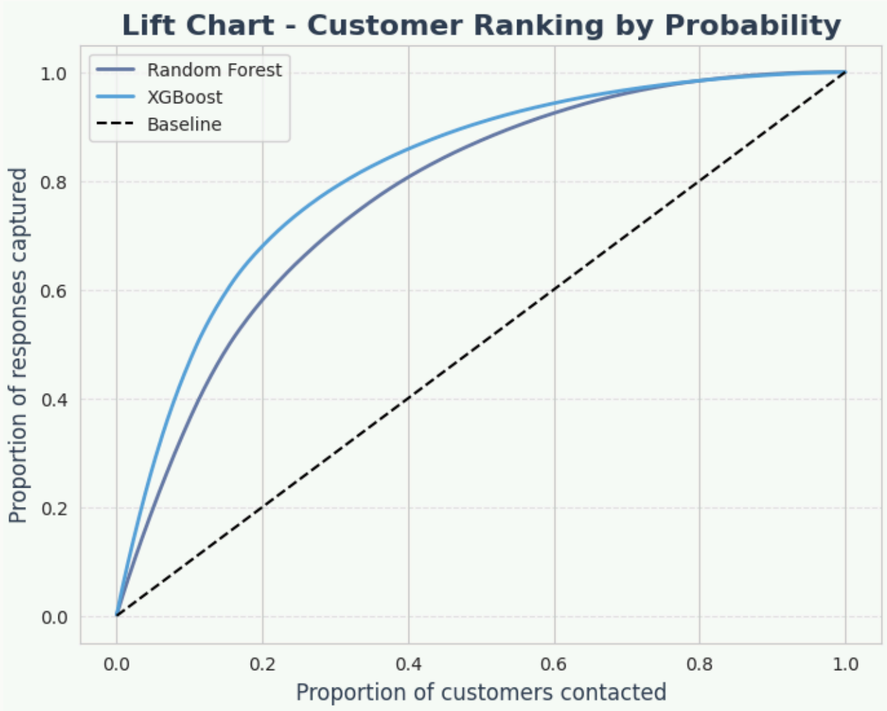
- The ROC curve comparison shows that both the Random Forest and XGBoost models perform well, with AUC values of 0.86 and 0.87, respectively. A high AUC indicates that the models are effective in distinguishing between buyers and non-buyers, which is crucial for optimizing marketing expenses. The slightly better AUC of XGBoost suggests it might offer a marginally better balance between detecting true buyers and minimizing false positives, potentially leading to better cost efficiency in targeting high-value customers while reducing wasted resources.

- The Precision-Recall curve comparison highlights differences in handling imbalanced data. The Random Forest model maintains higher precision at various recall levels, meaning it makes fewer incorrect predictions of buyers, which helps in reducing unnecessary marketing expenses. However, XGBoost exhibits better stability across recall levels, which might indicate a more consistent ability to capture buyers even at lower confidence thresholds. From a business perspective, if maximizing conversions is the priority, XGBoost might be preferable. However, if avoiding marketing costs for non-buyers is the goal, Random Forest could be the better choice.

05

PROFITABILITY ANALYSIS: COMPARISON OF THE TWO MODELS USING THE PROFIT CURVE

Ranking of customers by probability of response And Lift Chart - Customer Ranking by Probability



This graph shows the models' ability to identify the most likely customers to respond, by comparing Random Forest, XGBoost, and a baseline (random model).
The goal is to evaluate how much a model improves the identification of the best customers compared to a random selection.

Model Performance:

- The higher a curve is above the baseline, the more effective the model is at quickly capturing the customers who will respond.
- XGBoost is above Random Forest, meaning it captures responses more quickly than Random Forest.

Optimal Contact Threshold:

- If a model quickly reaches a plateau, it suggests that beyond a certain percentage of customers contacted, the additional effort no longer brings significant gains.
- The company can then set an optimal threshold to maximize the impact of its marketing campaigns at a reduced cost.

05

PROFITABILITY ANALYSIS: COMPARISON OF THE TWO MODELS USING THE PROFIT CURVE

Probability Distribution for the Top 15% Best Customers And Top 10 of Clients (Random Forest and Xgboost)

The first graph represents the distribution of response probabilities for the best customers (the top 15% according to the models) from Random Forest (RF) and XGBoost (XGB). The goal is to observe how each model evaluates the response probability of the most likely customers to purchase.

Difference in Distributions:

- The Random Forest (RF) curve is more spread out than the XGBoost curve, meaning the model assigns probabilities with more variability.
- The XGBoost (XGB) curve is also more concentrated, indicating a model that is more confident in its estimates.

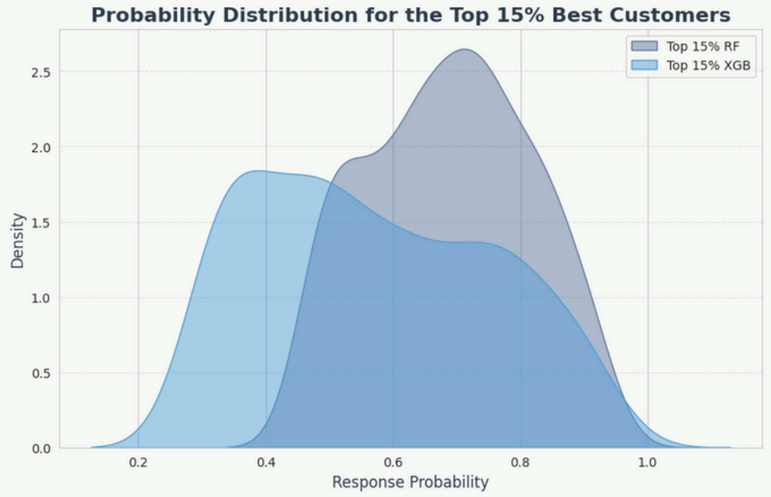
Comparison of Distribution Peaks:

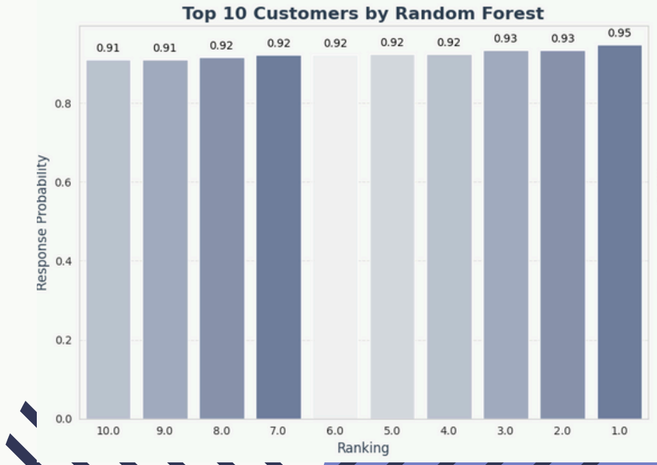
- Random Forest (RF) has a higher peak, meaning it assigns a more consistent probability to the majority of good customers.
- There is a longer tail to the right, potentially capturing customers with a very high response probability.

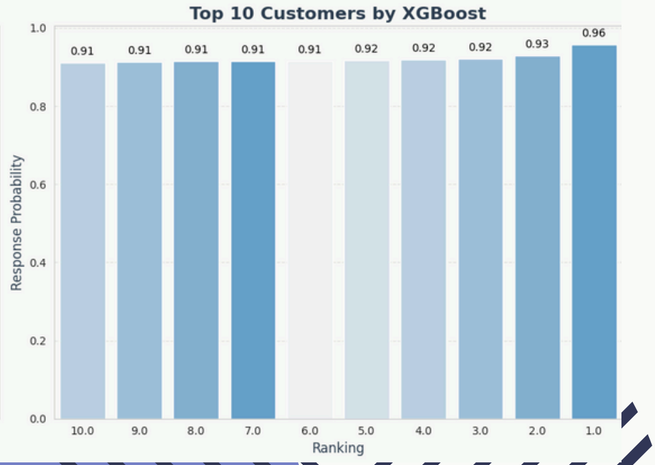
The second graph compare the top 10 customers identified by Random Forest (RF) and XGBoost (XGB) based on their response probability.

- Random Forest (RF): Probabilities range from 0.91 to 0.95, showing consistency.
- XGBoost (XGB): Probabilities range from 0.91 to 0.96, slightly more dispersed.

Key difference: XGBoost identifies one customer with a higher probability (0.96), while RF remains more stable.







05

PROFITABILITY ANALYSIS: COMPARISON OF THE TWO MODELS USING THE PROFIT CURVE

Profit Curve Analysis for Random Forest (RF) vs. XGBoost (XGB)

What is the best percentage of customers to be contacted to maximize profits ?

01

The graph represents the cumulative profit curve based on the number of clients contacted for two machine learning models: Random Forest (RF) and XGBoost (XGB).

General Trend:

- Profit curves initially increase, reach a maximum and then decrease. This indicates that beyond a certain threshold, it is no longer profitable to contact more customers.
- Declines after maximums mean that the remaining customers are generating a negative profit, probably due to the cost associated with contacting them.

Model Comparison:

- Random Forest (RF) reaches a higher maximum cumulative profit than XGBoost.
- XGBoost (XGB) reaches its maximum profit earlier, but with a lower profit.

02

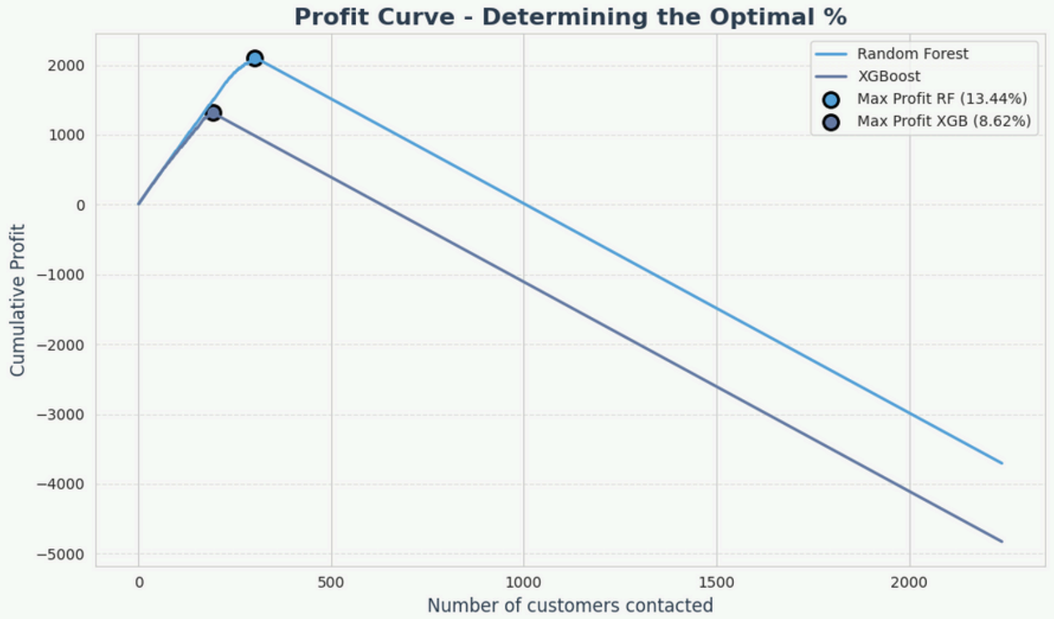
What is the best percentage of customers to be contacted to maximize profits ?

- **RF: 13.44% of 2240 → ~ the first 301 clients (ranked by probability of response)**
- **XGB: 8.62% of 2240 → ~ the first 193 clients (ranked by probability of response)**

These points are marked on the graph with black circles.

Interpretation:

- The RF model appears to be more profitable than XGBoost for maximizing overall profit.
- It's crucial to stop at the right point (around 301 clients for RF and 193 clients for XGB) to avoid losses.
- In a resource optimization context, RF would be more suitable if the goal is to maximize the gain with a larger margin.



06 CONCLUSION AND RECOMMENDATIONS

Summary

- Profitability: Random Forest (RF) generates a higher total profit (2100.00) with 13.44% of customers, compared to 1313.00 for XGBoost (XGB) with 8.62%.
- Efficiency: RF requires contacting more customers but yields higher returns, while XGB achieves its optimal profit with fewer contacts.
- Negative Profits: When all customers are contacted, both models generate losses, but RF's loss (-3706.00) is smaller than XGB's (-4828.00), confirming RF's overall profitability.

Conclusion: RF is more profitable but requires contacting more customers. The choice depends on the balance between maximizing profit and minimizing outreach costs.

Recommendations:

- If the goal is to maximize total profit, Random Forest (RF) is the better choice, as it generates higher returns despite requiring a larger customer outreach.
- If the goal is to minimize customer contacts while maintaining good profitability, XGBoost (XGB) is preferable, as it reaches its optimal profit with fewer contacts, making it more resource-efficient.
- Avoid contacting all customers, as both models result in negative profits beyond their optimal threshold. A well-defined stopping rule should be implemented to prevent unnecessary costs.
- Hybrid approach: Consider combining both models —using XGB for initial high-confidence targeting, followed by RF for broader, high-potential selections— to balance efficiency and profitability.

THANK YOU

Oumou SOW
Lath ESSOH
Tsiba RAZAFINDRAKOTO

M2 DS2E
Unistra