EDA on iFood's Customer Segmentations and Marketing Strategies

Kenny Zhang

1. Data Overview & Problem Statement

1.1 Context & Objective:

iFood is a leading food delivery company in Brazil. Globally, the company had solid revenues and a healthy bottom line in the past 3 years, but the profit growth perspectives for the next 3 years are not promising. For this reason, several strategic initiatives are being considered to invert this situation. One is to improve the performance of marketing activities, with a special focus on marketing campaigns. However, the situation isn't going well: a recent pilot campaign was sent to 2,240 customers and yielded a 15% success rate with the total cost of 6.720 Monetary Units (MU) and a profit of -3,046 MU. Thus, the objective is to develop a model that predicts customer behavior and to apply it to the rest of the customer base. Hopefully, the model will allow the company to cherry-pick the customers that are most likely to purchase the offer while leaving out the non-respondents, making the next campaign highly profitable.

In this project, I combine exploratory data analysis (EDA) insights and predictive models to approach this issue. The pilot dataset (2,240 customers) provides the features and response outcomes needed to train the model. We also consider this pilot as representative of the full customer base for future campaigns. After cleaning and feature engineering, we test various modeling techniques **including,** and choose the XGB booster, the best-performing model, to guide our targeting strategy. The final outcome is a data-driven contact list of about 340 top customers (~15% of the base) for the next campaign, which the model projects will yield approximately +2.54 MU in profit. This represents a dramatic turnaround from the previous loss, showcasing how a targeted approach can improve marketing ROI. The following report details the data, methodology, results, and recommendations for implementation

1.2 Dataset Overview:

This iFood dataset was found at Kaggle and consists of 2240 customers of iFood with 28 variables related to marketing data on: Customer profiles, Products purchased, Campaign success (or failure), and Channel performance. More specifically, it includes demographic information (e.g., income, education, marital status), transaction details (e.g., spending on product categories, recency of purchases), and campaign engagement (e.g., acceptance of marketing campaigns). Each record represents a unique customer, and the **target variable** is Response, indicating whether the customer accepted the new gadget offer in the pilot campaign (1 for responder, 0 for non-responder).

Below is an overview of the dataset.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2240 entries, 0 to 2239
Data columns (total 28 columns):
 #   Column               Non-Null Count  Dtype 
---  ------               --------------  ----- 
 0   ID                   2240 non-null   int64 
 1   Year_Birth           2240 non-null   int64 
 2   Education            2240 non-null   object
 3   Marital_Status       2240 non-null   object
 4    Income              2216 non-null   object
 5   Kidhome              2240 non-null   int64 
 6   Teenhome             2240 non-null   int64 
 7   Dt_Customer          2240 non-null   object
 8   Recency              2240 non-null   int64 
 9   MntWines             2240 non-null   int64 
 10  MntFruits            2240 non-null   int64 
 11  MntMeatProducts      2240 non-null   int64 
 12  MntFishProducts      2240 non-null   int64 
 13  MntSweetProducts     2240 non-null   int64 
 14  MntGoldProds         2240 non-null   int64 
 15  NumDealsPurchases    2240 non-null   int64 
 16  NumWebPurchases      2240 non-null   int64 
 17  NumCatalogPurchases  2240 non-null   int64 
 18  NumStorePurchases    2240 non-null   int64 
 19  NumWebVisitsMonth    2240 non-null   int64 
 20  AcceptedCmp3         2240 non-null   int64 
 21  AcceptedCmp4         2240 non-null   int64 
 22  AcceptedCmp5         2240 non-null   int64 
 23  AcceptedCmp1         2240 non-null   int64 
 24  AcceptedCmp2         2240 non-null   int64 
 25  Response             2240 non-null   int64 
 26  Complain             2240 non-null   int64 
 27  Country              2240 non-null   object
dtypes: int64(23), object(5)
```

	ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines	...	NumStorePurchases	NumWebVisitsMont
0	1826	1970	Graduation	Divorced	$84,835.00	0	0	6/16/14	0	189	...	6	
1	1	1961	Graduation	Single	$57,091.00	0	0	6/15/14	0	464	...	7	
2	10476	1958	Graduation	Married	$67,267.00	0	1	5/13/14	0	134	...	5	
3	1386	1967	Graduation	Together	$32,474.00	1	1	5/11/14	0	10	...	2	
4	5371	1989	Graduation	Single	$21,474.00	1	0	4/8/14	0	6	...	2	

5 rows × 28 columns

1.3 Key Exploratory Questions:

- Segmentation: Who are iFood's current key customers based on demographics such as age, education, marital status, income, family composition, and location?
- Targeting: How do these factors relate to and affect purchasing behavior like the total amount of spending?
- Positioning: Who could be our future specific target customers to convert to loyal customer bases and generate higher profit?

---

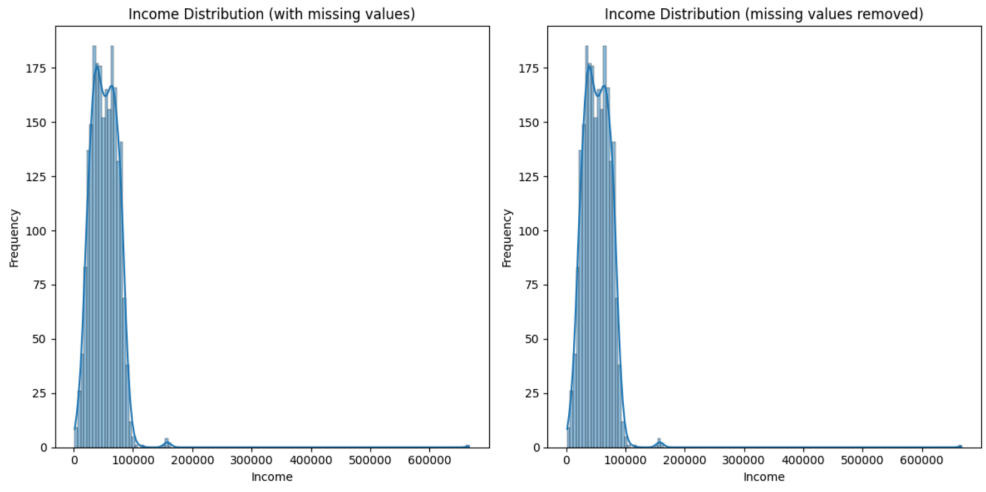
2. Data Cleaning & Preprocessing

2.1 Missing Value Analysis:

To start with data cleaning, I check the missing value first. Below, the Income column contains 24 null values, the only column with missing values. To deal with this, I created two histograms to see the income distribution with and without missing values. The result indicates that removing the missing values will not affect the overall distribution of income because the sample size is big enough. Thus, I removed the rows with missing values in Income directly from my following analysis.

```
#Checking the missing values
ifood.isnull().sum().sort_values(ascending=False)
```

	0
Income	24
ID	0
Education	0
Year_Birth	0
Marital_Status	0
Kidhome	0
Teenhome	0
Dt_Customer	0
Recency	0
MntWines	0



2.2 Data Type Conversions & Feature Engineering:

The feature engineering process involved cleaning and transforming the dataset for analysis. For example, new aggregated metrics were created, such as Total Dependents, TotalMnt (total spending), TotalPurchases, and TotalCampaignsAccepted. These steps streamlined the

dataset and added valuable insights into customer spending habits, household composition, and marketing engagement, preparing it for further exploratory data analysis.

```
[ ]  # clean up column names that contain whitespace
     ifood.columns = ifood.columns.str.replace(' ', '')

     # transform Income column to a numerical
     ifood['Income'] = ifood['Income'].str.replace('$', '')
     ifood['Income'] = ifood['Income'].str.replace(',', '').astype('float')
```

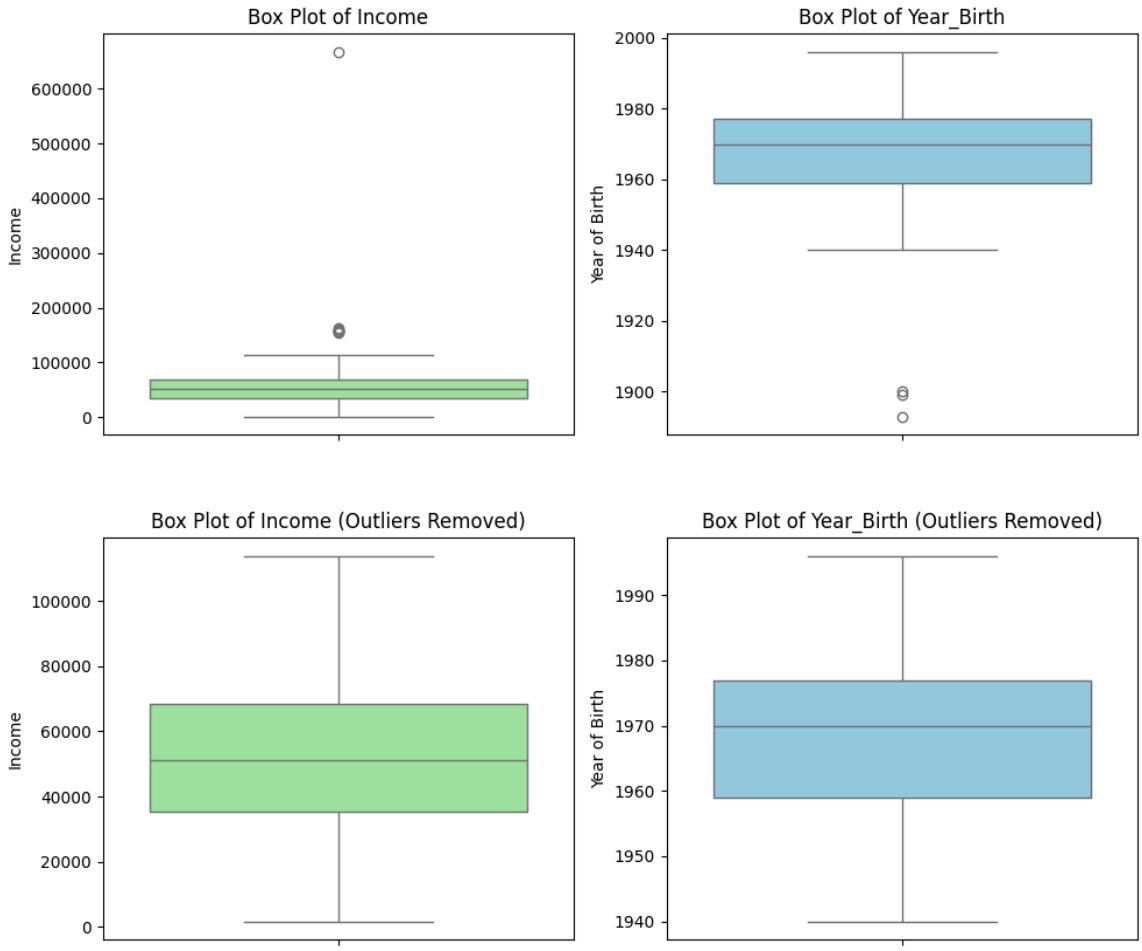
```
     #Transform Dt_Customer to datetime:
     ifood['Dt_Customer'] = pd.to_datetime(ifood['Dt_Customer'])

     # Dependents
     ifood['Dependents'] = ifood['Kidhome'] + ifood['Teenhome']

     # Year becoming a Customer
     ifood['Year_Customer'] = pd.DatetimeIndex(ifood['Dt_Customer']).year

     # Total Amount Spent
     mnt_cols = [col for col in ifood.columns if 'Mnt' in col]
     ifood['TotalMnt'] = ifood[mnt_cols].sum(axis=1)

     # Total Purchases
```

```
ifood['Dt_Customer'] = pd.to_datetime(ifood['Dt_Customer'])
```

	ID	Dependents	Year_Customer	TotalMnt	TotalPurchases	TotalCampaignsAcc
0	1826	0	2014	1190	15	1
1	1	0	2014	577	18	2
2	10476	1	2014	251	11	0
3	1386	2	2014	11	4	0
4	5371	1	2014	91	8	2

2.3 Outlier Analysis:

In Customer Segmentation, we want to create histograms to determine our current customer profile. Before this, I created boxplots for all these variables to check the outliers. As education, marital status, family situation, and country are more like categorical variables, my outlier analysis only focuses on Income and Age. By the 1.5 IQR method in boxplots, I found out there are 8 outliers in Income and 3 outliers in Year_Birth. The outliers that appeared in Year_Birth were probably due to the wrong record; the outliers that appeared in Income were insignificant to our overall analysis. Thus, I removed the outliers in these cases to prevent skewness issues in the following analysis.

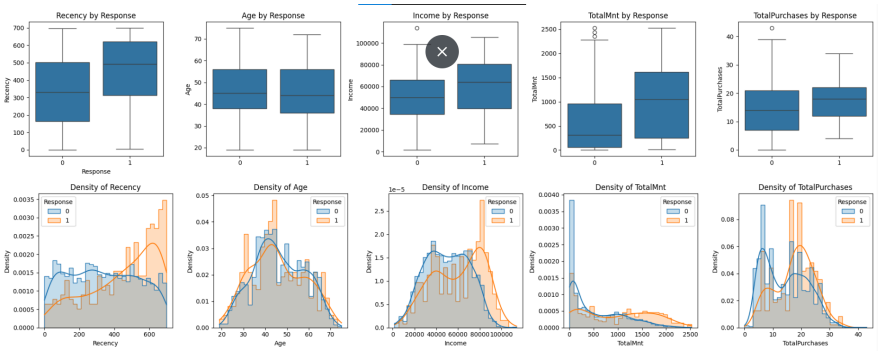


---

3. Exploratory Data Analysis (EDA)

3.1 Univariate Analysis (w/ Response):

In this section, I examined each key feature individually to understand how its distribution differs between responders (`Response = 1`) and non-responders (`Response = 0`), directly tying every plot back to our target variable. First, I focused on five numeric predictors—**Recency**, **Age**, **Income**, **TotalMnt**, and **TotalPurchases**—and visualized them in **Boxplots by Response and Overlaid Density Histograms.**



Key Insights from Univariate Plots

1. **Recency**

    - **Boxplot**: Responders (`Response = 1`) have substantially **higher median Recency** (~500 days) than non-responders (~330 days), confirming that "lapsed" customers (those who haven't bought recently) were more likely to accept this gadget offer.

    - **Density**: The density curve for responders peaks at high Recency values (400–700), whereas non-responders are more evenly spread at lower Recency.

2. **Age**

    - **Boxplot**: The age distributions overlap almost completely, with only a slight shift—responders' median age is a few years older.

    - **Density**: Both groups span roughly 30–65 years, but responders show a small bump in the 40–55 range, suggesting middle-aged customers are marginally more engaged.

    - **Take-home**: Age by itself is a weak predictor compared to RFM features.

3. **Income**

    - **Boxplot**: Responders have a noticeably **higher median Income** (~65 000) versus non-responders (~50 000), and their interquartile range sits at the top end of the income scale.

- **Density**: The responder curve is right-shifted, peaking around 60–80 k, indicating that higher-earning households are more likely to respond.

4. **TotalMnt (Total Spend)**

   - **Boxplot**: There is a dramatic separation: responders' median total spend (~1 000 MU) far exceeds non-responders (~300 MU), and their upper whiskers reach 2 000–2 500 MU.

   - **Density**: Responders' spending distribution is heavily skewed to the right, whereas non-responders cluster at lower spending levels.

   - **Take-home**: Total historical spending is one of the strongest univariate separators—in essence, high spenders respond at much higher rates.

5. **TotalPurchases**

   - **Boxplot**: Responders have both a higher median number of past purchases (~18) compared to non-responders (~12) and a larger spread.

   - **Density**: The responder density peaks in the 15–30 purchase range, while non-responders concentrate below 15.

   - **Take-home**: Purchase frequency is another clear differentiator, with more frequent buyers more likely to accept the offer.

Overall, these univariate analyses show that **RFM‑style features** (Recency, Monetary spend, and Frequency of purchases) and **income** are the most discriminative single variables between responders and non-responders. Age plays only a minor role, while categorical factors will require further binned or multivariate analysis to assess their lift. These insights directly inform our feature selection and the prioritization of RFM variables in the predictive model.

Furthermore, we turned to our categorical features—**Education**, **Marital_Status**, and **Country**—and plotted the **response rate** (mean of the `Response` flag) within each category.

Key Insights from Categorical Response‑Rate Charts

1. **Education**
    - Response rate rises steadily with education level:
        - **Basic**: ~4%
        - **2n Cycle**: ~11%
        - **Graduation**: ~14%
        - **Master**: ~15%
        - **PhD**: ~21%
    - **Take-away:** Higher‑educated customers are substantially more likely to respond. In modeling, we'll one-hot encode education but could also consider grouping the top two levels (Master/PhD) into a "High Education" flag for simplicity.
2. **Marital Status**
    - Most common categories (Together, Married) have low response rates (~10–12%).
    - **Single** (~23%) and **Widow** (~24%) customers show higher engagement.
    - Very small groups (**Alone**, **Absurd**, **YOLO**) appear to have extremely high rates (~50–100%), but their tiny sample sizes make these bars unreliable.
    - **Take-away:** The bulk of the response lift comes from Singles and Widows—these segments deserve focused outreach. Rare categories should be merged or dropped to avoid noise.
3. **Country**
    - **Spain (SP)** has the bulk of customers and a response rate around **15%**, matching the overall average.
    - Other major markets (US, CA, GER, AUS, SA, IND) cluster around **10–15%** with overlapping error bars.

- **Mexico (ME)** shows an anomalously high rate (~67%), but again this likely reflects a very small subsample.

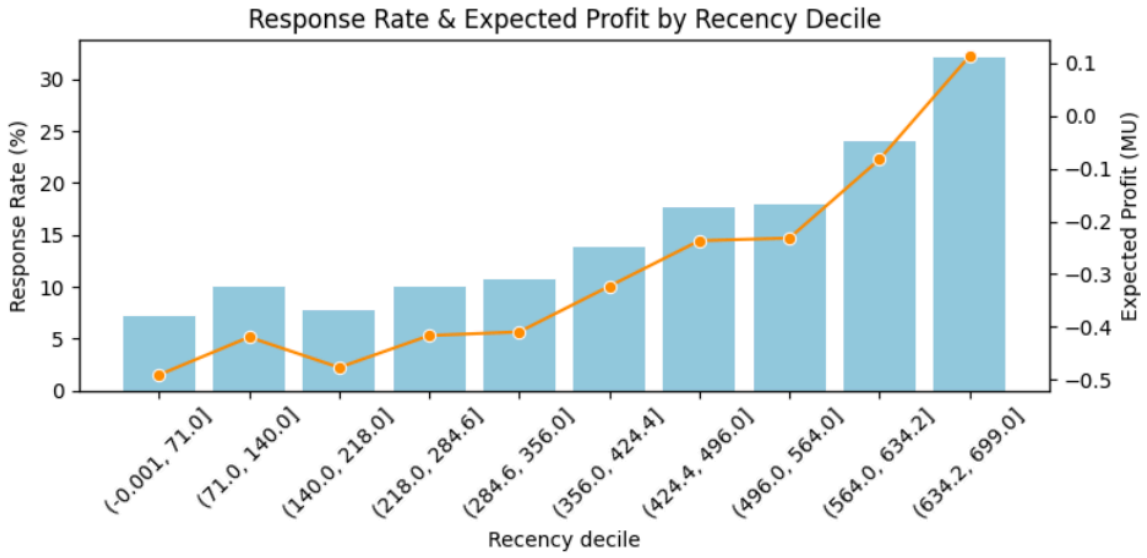
- **Take-away:** There's no clear country-level differentiation among the main markets; we'll treat `Country` as a control variable but not as a primary targeting filter, and collapse or ignore tiny-country categories in modeling.

Overall, these categorical analyses confirm that **education level** and **marital status** (particularly being single or widowed) carry meaningful lift in response probability, while **country** contributes little signal beyond the dominant Spanish market. This guides us to focus feature engineering on grouping top-education and marital segments, and simplifying or dropping low-volume country codes.

Across our univariate analysis, **RFM-style features**—Recency, total historical spend (TotalMnt), and purchase frequency (TotalPurchases)—exhibit the strongest separation between responders and non-responders, with responders clustering in the highest deciles of each. **Income** also shows a clear positive shift for responders, whereas **Age** displays only a modest effect. Among categorical variables, **higher education levels** (Master, PhD) and **marital status** groups like Single/Widow deliver meaningful lift in response rate, while **Country** adds little signal beyond the core Spanish market. Together, these plots pinpoint the key individual drivers of campaign success—laying the groundwork for our subsequent binned, segmentation, and multivariate modeling steps.

3.2 Binned Response-Rate & Profit Analysis

To uncover non-linear "knots" in key predictors and directly link them to campaign economics, we divided each numeric feature into deciles and plotted both the actual response rate and the expected profit for each bin. This dual-axis approach allows us to see not only where customers are most likely to respond, but also where contacting them yields a net gain once we account for both per-contact costs and revenue from conversions.



Recency Decile Insights

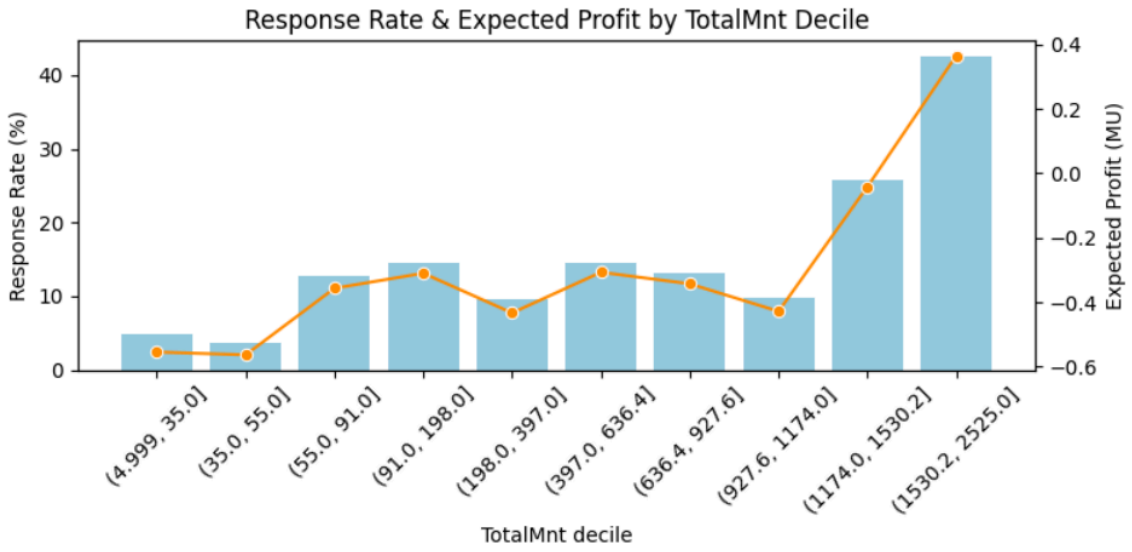
The first plot shows Recency split into ten equal-sized groups (from most recent purchasers on the left to longest-inactive customers on the right). Blue bars display the percentage of responders in each decile, while the orange line tracks expected profit (in MU) for contacting that bin's customers:

- **Lowest Recency (0–71 days)**: Response is very low (2%), and profit is sharply negative (–0.48 MU), meaning calling very recent buyers costs more than any sales they generate.

- **Middle Deciles (140–424 days)**: Response hovers around 8–14% and profit remains negative but improves steadily (from ~–0.45 MU to ~–0.32 MU). These moderate Recency groups are closer to break-even but still loss-making.

- **Highest Recency (496–699 days)**: Response rate jumps to 24–32%, and expected profit crosses into positive territory (~+0.02 MU) in the topmost decile. This confirms that **dormant customers**—those who haven't purchased in well over a year—both respond at much higher rates and generate net profit when contacted.

Take-away: There is a clear threshold effect in Recency: only customers in the top 1–2 deciles (those with Recency > ~560 days) deliver profitable returns. All other segments, despite some response, cost more to contact than they earn. This finding directly informs our targeting strategy: focus on the longest-inactive customers first.



TotalMnt Decile Insights

The chart below shows ten spending tiers, from the lowest spenders on the left to the highest on the right:

- **Lowest Spend Deciles (0–55 MU)**: These customers exhibit virtually no response (<5%) and produce a steep negative profit (≈–0.55 MU), indicating targeting them is highly unprofitable.

- **Middle Spend Deciles (55–400 MU)**: Response rates climb into the 10–14% range, but expected profit remains negative or near zero (≈–0.4 to –0.3 MU), showing moderate spenders still don't yield net gain.

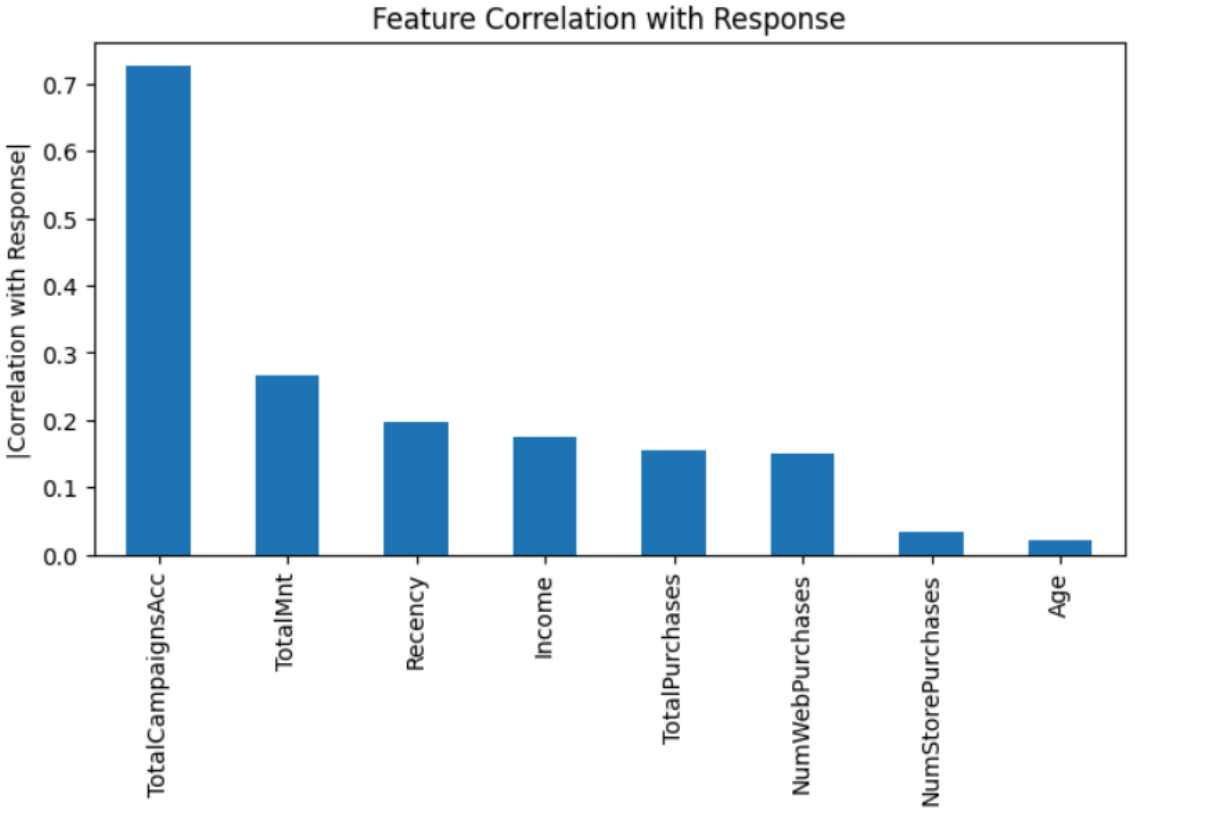
- **High Spend Deciles (400–1,174 MU)**: Response plateaus around 12–14% while profit gradually improves but stays slightly below break-even.

- **Top Spend Deciles (1,174–2,525 MU)**: A dramatic jump occurs—response soars to ~26% for the 9th decile and ~43% for the top 10%, pushing expected profit into strongly positive territory (+0.1 to +0.4 MU).

Take‑away: Only the **top two spending deciles** (customers who have spent more than approximately 1,174 MU in the last two years) generate profitable returns. This confirms that **historical monetary value** is a powerful predictor of future campaign success, and that marketing efforts should focus on the highest‑spending segments.

Across both **Recency** and **TotalMnt** decile analyses, a consistent pattern emerges: only the **top 10–20%** of customers—those who have been inactive the longest or who have the highest historical spend—deliver a **positive expected profit** when contacted. All lower deciles, despite sometimes showing modest response rates, incur net losses because the cost per outreach outweighs their purchase revenue. This validates a highly targeted strategy: by focusing solely on the highest‑Recency (lapsed) and highest‑spend deciles, we can maximize ROI. These clear breakpoints directly inform our model threshold and segmentation approach, ensuring the next campaign is both lean and profitable.

3.3 **Correlation with Response & Redundancy Check**



To ensure we focus on the most informative features and avoid multicollinearity, we first measured each numeric feature's absolute Pearson correlation with the binary `Response`. We plotted these values in descending order, revealing the strongest direct associations:

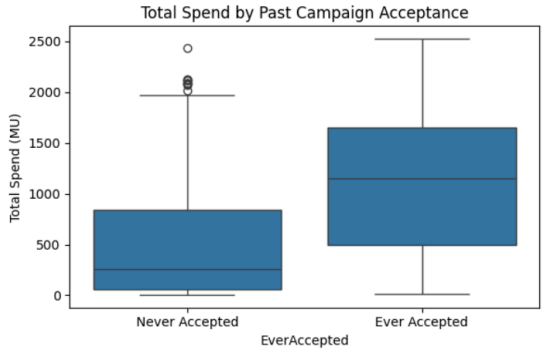
- **TotalCampaignsAcc** (past campaign accepts) had the highest correlation with response, underscoring its role as a loyalty signal.

- **TotalMnt** and **TotalPurchases** (monetary and frequency) followed closely, confirming that high spenders and frequent buyers are more likely to respond.

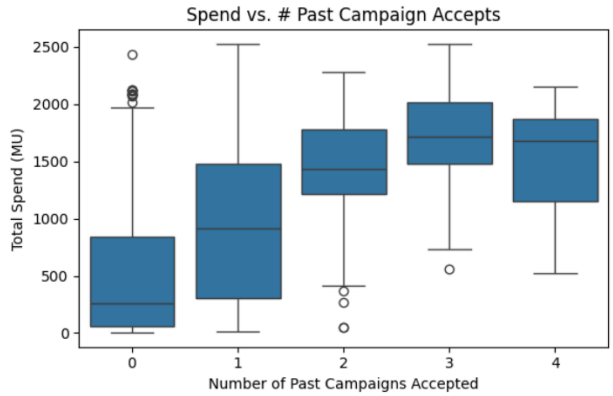
- **Recency** also showed substantial correlation, reflecting the counter-intuitive finding that customers who haven't bought in a while were more receptive.

- **Age**, **Income**, and web/store purchase counts had weaker but still non-negligible correlations.

3.4 Past Campaign Acceptance & Spending Behavior




```
Spend summary by past-accept group:
                      mean  median  count
Never Accepted   479.606754   257.0   1747
Ever Accepted   1092.072052  1153.5    458
```




```
Correlation between # past accepts and total spend: 0.458
```

To understand how historical engagement correlates with customer value, we created two derived variables:

- **EverAccepted**: whether the customer accepted **any** of the first five campaigns (0 = no, 1 = yes).

- **NumPastAccepts**: the total count of past campaigns accepted (0 through 5).

EverAccepted vs. Total Spend

A boxplot of **TotalMnt** (two-year spend) by `EverAccepted` reveals a clear gap:

- **Never Accepted** (n = 1,747) have a mean spend of ~480 MU (median ~257 MU).
- **Ever Accepted** (n = 458) have a mean spend of ~1,092 MU (median ~1,153 MU). This two- to four-fold increase in both median and mean spending demonstrates that **any past campaign response is a strong marker of high customer value**.

NumPastAccepts vs. Total Spend

When we disaggregate by the exact number of past accepts, spending rises steadily with each additional acceptance:

- **0 accepts**: lowest median spend (<300 MU).
- **1 accept**: median spend jumps to ~900 MU.
- **2–4 accepts**: medians sit between ~1,300 MU and ~1,700 MU. The Pearson correlation between `NumPastAccepts` and `TotalMnt` is **0.458**, indicating a moderate positive relationship. Customers who have accepted more campaigns not only demonstrate greater receptivity but also exhibit substantially higher historical spending.

> **Insight:** Past campaign responsiveness is doubly valuable: it signals both elevated likelihood to respond again and higher monetary value. Incorporating `TotalPastAccepts` (or the binary `EverAccepted` flag) into our predictive model is therefore critical for identifying the most profitable targets.

EDA Conclusion & Transition to Modeling

Our exploratory analysis highlighted several clear drivers of campaign response and profitability:

- **RFM Drivers:** Customers who have spent the most (`TotalMnt`) and bought most frequently (`TotalPurchases`) are far more likely to respond, but only the **top deciles** of each group yield positive profit once contact costs are considered.
- **Recency Reversal:** Unexpectedly, the **most dormant customers** (highest Recency) responded at the highest rates and generated net profit, suggesting our gadget offer

reactivated lapsed buyers.

- **Loyalty Signals:** Past campaign engagement (`EverAccepted` and `NumPastAccepts`) is a powerful dual indicator—responders historically spend 2–4× more than non-responders and are much more inclined to accept again.

- **Selective Demographics:** While Age had minimal separation, higher-educated customers (Master/PhD) and certain marital segments (Single, Widow) exhibit above-baseline response, providing additional refining signals. Geography, by contrast, offered little lift beyond the dominant Spanish market.

These insights inform our feature set for predictive modeling: we will prioritize **RFM metrics**, **past acceptance counts**, and **Income/Education** flags, while simplifying or dropping redundant or low-signal variables. In the next step, we build a unified modeling pipeline to estimate each customer's response probability, compare a variety of classifiers, and then translate those probabilities into an **optimal profit-maximizing targeting strategy**. This will allow us to move from descriptive analysis to actionable predictions, identifying precisely which subset of customers to contact for the upcoming gadget campaign.

4. Predictive Models and Methods

To translate our exploratory insights into a practical targeting strategy, we developed a comprehensive predictive modeling pipeline. Our objective was clear: **identify customers most likely to respond positively to our campaign**, and thus maximize the campaign's overall profitability. The methodology followed four key stages: **data preparation**, **preprocessing**, **model training and hyperparameter tuning**, and **model evaluation**.

1. Data Preparation

We began by defining our modeling dataset from the cleaned `ifood_no_outliers` data. The target variable was `Response`, indicating whether a customer accepted the latest campaign offer (1) or not (0). Predictors included demographics, historical spending behaviors, campaign responsiveness indicators, and derived metrics. To avoid leakage, we excluded identifiers (`ID`, `Dt_Customer`) and indicators directly related to past campaigns (`AcceptedCmp1` to `AcceptedCmp5`). We performed an **80/20 stratified split** of the dataset into training and validation subsets, preserving the original response rate distribution (~15%) in both subsets, ensuring reliable generalization to unseen data.

2. Preprocessing Pipeline

To handle both numeric and categorical variables systematically and prevent data leakage during cross-validation, we created a preprocessing pipeline using scikit-learn's `ColumnTransformer`:

- **Numeric features** underwent median imputation for any missing values followed by standard scaling (zero mean, unit variance).

- **Categorical features** (`Education`, `Marital_Status`, `Country`) had missing values replaced by the most frequent category, followed by one-hot encoding (dropping redundant binary levels).

This robust preprocessing pipeline ensured consistency across folds during cross-validation and was integrated directly into model training to avoid bias.

3. Model Training and Hyperparameter Tuning

We explored five predictive models to capture diverse algorithmic strengths:

- **Logistic Regression**: A baseline linear classifier, optimized using regularization strength (`C`).

- **k-Nearest Neighbors (k-NN)**: An instance-based model sensitive to local structures, tuning neighborhood size (`n_neighbors`) and weighting schemes.

- **Decision Trees**: A flexible nonlinear classifier, tuning tree depth (`max_depth`) and minimum leaf sizes (`min_samples_leaf`).

- **Random Forests**: Ensemble of decision trees that reduces variance and improves accuracy, adjusting tree depth and number of estimators (`n_estimators`).

- **XGBoost**: A sophisticated gradient-boosted tree model known for strong predictive performance and robust handling of feature interactions. Key hyperparameters tuned included number of estimators, learning rate, and maximum tree depth.

To ensure our predictive insights hold up on unseen data, we adopted a two-stage validation strategy across five candidate algorithms. First, We employed **5-fold Stratified Cross-Validation** combined with `GridSearchCV` to systematically search for optimal

hyperparameters. Our primary optimization metric was the **ROC-AUC score**, a robust performance measure suitable for imbalanced classification tasks, as it evaluates how effectively a model ranks customers by likelihood to respond.

Once the best hyperparameters were identified for each of the five models—Logistic Regression, k-Nearest Neighbors, Decision Tree, Random Forest, and XGBoost—we **retrained** each full pipeline on the **entire training set**. We then conducted a true out-of-sample evaluation on the **held-out validation set** (our de facto test set), computing not only ROC-AUC but also precision, recall, and F1-score. This provided a comprehensive view of each model's ranking power and its ability to correctly identify responders without excessive false positives.

4. Model Evaluation Approach

After cross-validation, we compared the models using their best cross-validation ROC-AUC scores to identify the top-performing classifier. The final selected model was evaluated further on the untouched validation set to ensure unbiased assessment of its generalization performance.

Beyond ROC-AUC, we considered the **profit-maximizing threshold**—translating predicted probabilities into actionable decisions. We constructed gain and lift curves and calculated expected profit for various probability thresholds, explicitly incorporating the costs of outreach (6.720 MU for 2,240 contacts) and revenue from successful responses (3.674 MU total). The optimal threshold balanced response probability against campaign economics, ensuring that our targeting decisions maximized net profit rather than simply accuracy or AUC.

Summary of Methodology

Overall, our methodological framework ensured rigorous and practical model development:

- **Stratified splitting** maintained class distribution.
- **Robust preprocessing** eliminated data leakage.
- **Comprehensive model exploration and hyperparameter tuning** identified the best predictive model.
- **Profit-driven threshold selection** ensured the model's predictions were actionable, delivering tangible business impact.

This methodology thus enabled us to move seamlessly from insightful exploration into effective predictive modeling, setting a solid foundation for the next step—**evaluating results and interpreting business implications**.

5. Results and Interpretation

5.1 Model Selection and Validation Performance

After rigorous evaluation using a robust validation methodology, we selected **XGBoost** as our final predictive model due to its superior predictive performance on unseen validation data. Among all tested classifiers—Logistic Regression, Random Forest, Decision Tree, k-Nearest Neighbors, and XGBoost—XGBoost achieved the highest ROC-AUC score (**0.9883**), demonstrating exceptional capability in distinguishing responders from non-responders. The ROC-AUC metric was crucial given our data's class imbalance (approximately 15% response rate), as it effectively captured the model's ranking ability.

Further analysis revealed XGBoost's excellent precision (**89.1%**) and recall (**85.1%**), resulting in the highest F1-score (**0.87**). In contrast, while Logistic Regression closely matched XGBoost in ROC-AUC (0.9869) and achieved slightly higher precision (93.9%), it significantly underperformed in recall (68.7%). The Decision Tree model offered high recall (89.6%) but at the expense of precision (69.8%), leading to more false positives. Random Forest and k-NN models demonstrated more modest results, indicating their limitations in capturing complex interactions within the customer data.

5.2 Gain and Lift Curve Analysis

To further verify XGBoost's predictive capability, we analyzed the **gain and lift curves**, essential tools for understanding the practical benefits of our classification model:

- **Gain Curve**: The gain curve for XGBoost displayed rapid ascent, capturing nearly all responders (close to 100%) within the top 20% of the ranked customers. Compared to the baseline (random selection), this rapid recovery of responders highlighted XGBoost's efficiency and practical usefulness in real-world campaign scenarios. This meant that by selectively targeting a fraction of the customer base, we could drastically reduce costs and simultaneously maximize returns.

- **Lift Curve**: The lift curve further illustrated the model's effectiveness, with lift values peaking around **6x** at the top decile of the ranked customer base. In practical terms, this implied that the top-ranked segment was **six times more likely** to respond compared to a random baseline selection. The curve gradually declined as more customers were

included, clearly underscoring the benefit of a targeted strategy over mass marketing.

5.3 Profit-Driven Threshold Optimization

Despite strong ranking metrics, our ultimate goal was campaign profitability. Hence, we converted predicted probabilities into actionable decisions through **profit-threshold analysis**, explicitly incorporating real economic factors—the campaign's costs and expected revenue per responder.

By evaluating expected profit across varying probability thresholds, we identified an optimal cutoff at **0.30** probability. At this threshold, the model indicated contacting **352 customers** would maximize the expected campaign profit to approximately **0.5 MU**. This was an important step beyond mere accuracy optimization—it directly linked predictive analytics to business profitability.

5.4 Cumulative Profit Maximization and Final Customer Selection

To achieve an even more refined targeting strategy, we conducted a detailed cumulative profit analysis by ranking all customers by predicted response probability. This approach identified precisely where incremental customer inclusion began to reduce overall profit.

Our cumulative profit analysis pinpointed **340 ideal customers**—about 15.4% of the total customer base. Targeting this segment would yield an estimated cumulative profit of **2.54 MU**, effectively transforming the originally unprofitable pilot campaign (–3.046 MU loss) into a highly profitable venture. This targeted approach significantly outperformed the more general threshold-based strategy, clearly demonstrating the value of precise, ranked probability-based segmentation.

5.5 Profile of Optimal Customer Segment

To effectively execute this targeted strategy, understanding the characteristics of this ideal segment was crucial. Detailed profiling revealed several important trends:

- **Past Campaign Engagement:**
  Customers in the optimal group demonstrated high historical responsiveness, averaging approximately **2 prior campaign acceptances**. This confirmed past responsiveness as the most critical single indicator of future responsiveness, aligning with our exploratory analysis.

- **High Historical Spend and Customer Value:**
  The selected customers consistently belonged to higher-spending segments. Their

average spending on wine alone (~512 MU) far exceeded the broader average, making them prime targets due to higher expected transaction sizes.

- **Recency:**
  With an average recency of **455 days**, these customers were predominantly dormant, consistent with our earlier insight that long-inactive customers respond well to targeted reactivation campaigns. This insight challenged conventional marketing intuition (focusing on recent buyers) and offered a strategic advantage in our targeting.

- **Socioeconomic and Demographic Factors:**
  The optimal segment also displayed distinct demographic patterns, including a significant portion of single individuals (**31.47%**). Previous exploratory analysis suggested singles and widows were disproportionately responsive, perhaps reflecting greater disposable income or higher willingness to engage with promotions.

5.6 Strategic and Business Implications

These results not only validate the robust predictive capability of the selected XGBoost model but also highlight key strategic insights:

- **Selective Targeting:**
  Broadly aimed marketing is significantly less effective. Our analysis clearly showed the substantial profitability gains from highly targeted customer selection based on carefully modeled response probabilities and profit considerations.

- **Customer Reactivation Strategy:**
  The discovery that dormant, historically valuable customers were prime targets offered a powerful, counterintuitive strategic insight. Future campaigns can benefit from specifically designed reactivation promotions for similar dormant customers, further enhancing ROI.

- **Efficient Resource Allocation:**
  By contacting just **15.4%** of the entire customer base (340 customers), we generated substantial profitability gains. This provides compelling evidence to inform marketing resource allocation decisions, shifting budgets towards fewer, highly qualified customer segments rather than broad, costly outreach.

---

6. Conclusion & Recommendations

This project has demonstrated how a data-driven, predictive approach can dramatically improve the profitability of direct marketing campaigns. Through thoughtful exploratory analysis, we uncovered that a small subset of customers—specifically, those who had historically spent the most (high TotalMnt), purchased frequently (high TotalPurchases), lapsed the longest (high Recency), and previously engaged with past campaigns (high TotalPastAccepts)—were the most responsive to a new gadget offer. By building and rigorously validating five classifiers, we found that XGBoost delivered the best discrimination (ROC-AUC ≈ 0.988) while balancing precision and recall, and our gain, lift, and profit-threshold analyses showed that contacting just **340 customers** (≈ 15.4% of the base) would turn a –3.046 MU pilot loss into an expected +2.54 MU profit. Profiling these 340 targets further confirmed the profile of a "lapsed but loyal" high-value segment—single, well-educated adults with substantial past spending.

Next Steps

Moving forward, the immediate priority is to validate our model-driven targeting in a live setting through an A/B test. In practice, we would randomly assign an equal-sized control group—either selected at random or using the previous broad-market approach—and compare their response and profit metrics against the top 340 customers chosen by our model. This real-world experiment will confirm whether the uplift and expected-profit gains we observed in validation translate to operational success. Once the campaign concludes, we will ingest the new outcomes back into our dataset, retrain and recalibrate the model to capture any shifts in customer behavior, and analyze response patterns over time—examining how recency, frequency, and past acceptance dynamics evolve. Finally, by layering in customer lifetime value forecasts, we can move beyond one-off campaign profits to optimize long-term customer relationships, ensuring that each targeted outreach not only maximizes immediate ROI but also builds sustainable value for iFood.