Profit-Driven Customer Targeting through Predictive Analytics: An iFood Campaign Case Study

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. After cleaning and feature engineering, I test various classification models, including logistic regression, random forest, decision tree, KNN, and XGB. By comparing the performance metrics, I chose 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

---

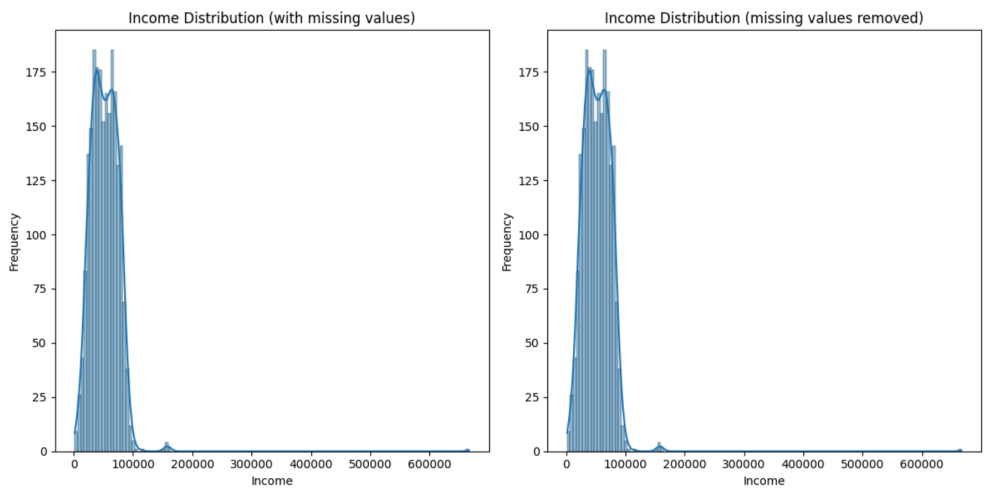
2. Data Cleaning & Preprocessing

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')
```

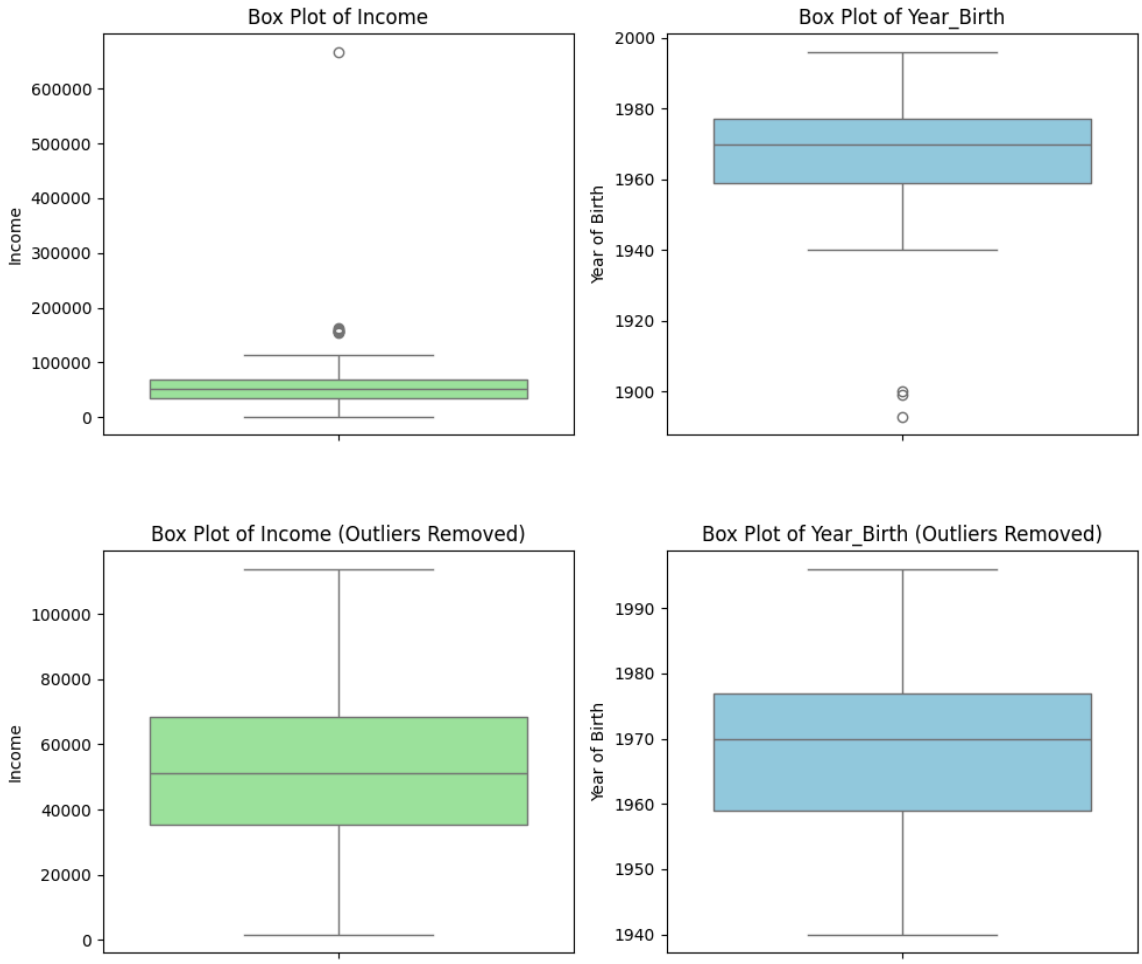
```
#Transform Dt_Customer to datetime:
ifood['Dt_Customer'] = pd.to_datetime(ifood['Dt_Customer'])

# 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:

I created box plots 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.

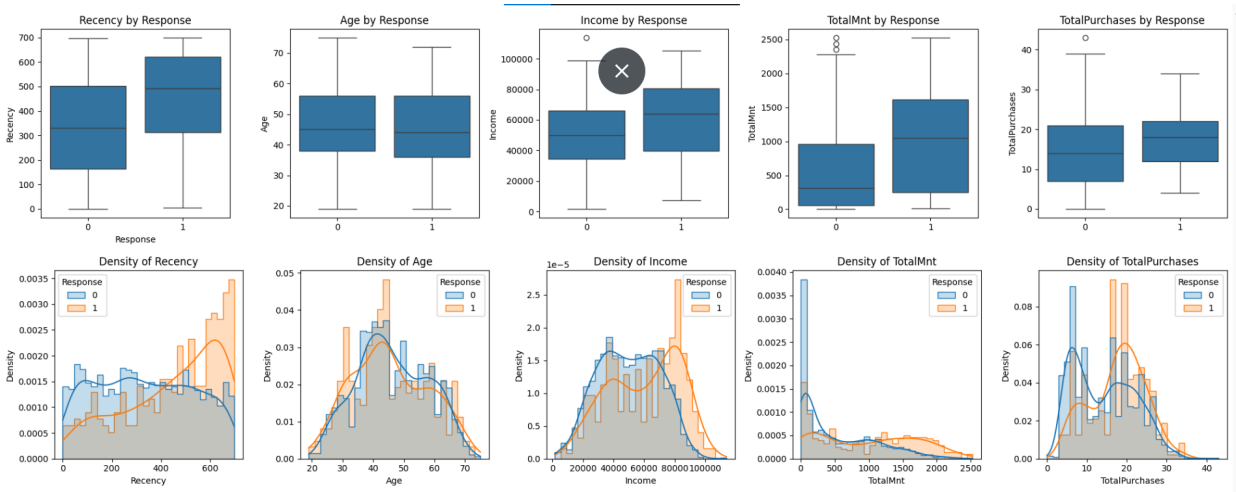


---

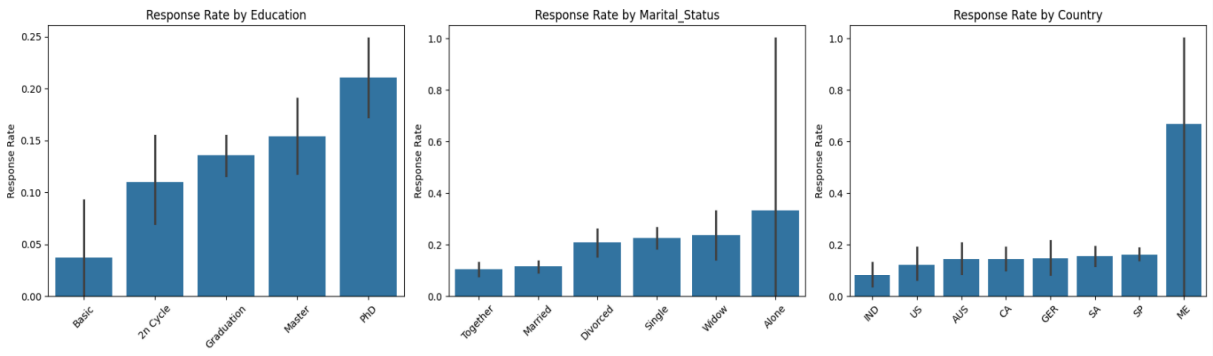
3. Exploratory Data Analysis (EDA)

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, including **Recency**, **Age**, **Income**, **TotalMnt**, and **TotalPurchases**, and visualized them in boxplots by Response and overlaid density histograms.



Across our univariate analysis, the classic Recency, Frequency, and Monetary value (RFM) metrics emerged as the strongest individual predictors of response. **Recency** showed that the most dormant customers (median ~500 days since last purchase) were far more likely to accept the gadget offer than recent buyers (~330 days), turning conventional intuition on its head. **Total historical spending (TotalMnt)** likewise offered dramatic separation—responders' median spend (~1,000 MU) dwarfed non-responders (~300 MU). **Purchase frequency (TotalPurchases)** echoed this pattern: high-frequency buyers (median ~18 past purchases) responded at much higher rates than those with fewer transactions (median ~12). **Income** also carried a right‑shift for responders (peaking around 60-80 K versus 50 K), whereas **Age** showed only a modest effect. These insights directly inform our feature selection and the prioritization of RFM variables in the further predictive model.

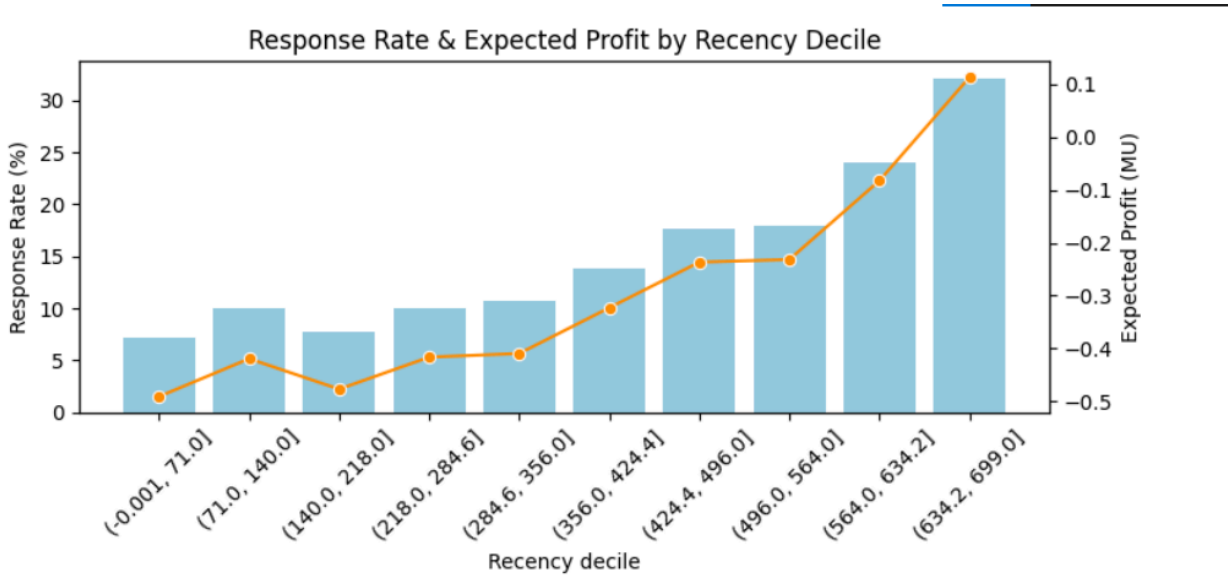


Moreover, I turned to our categorical features—**Education**, **Marital_Status**, and **Country**—and plotted the response rate within each category. Across our categorical analysis,

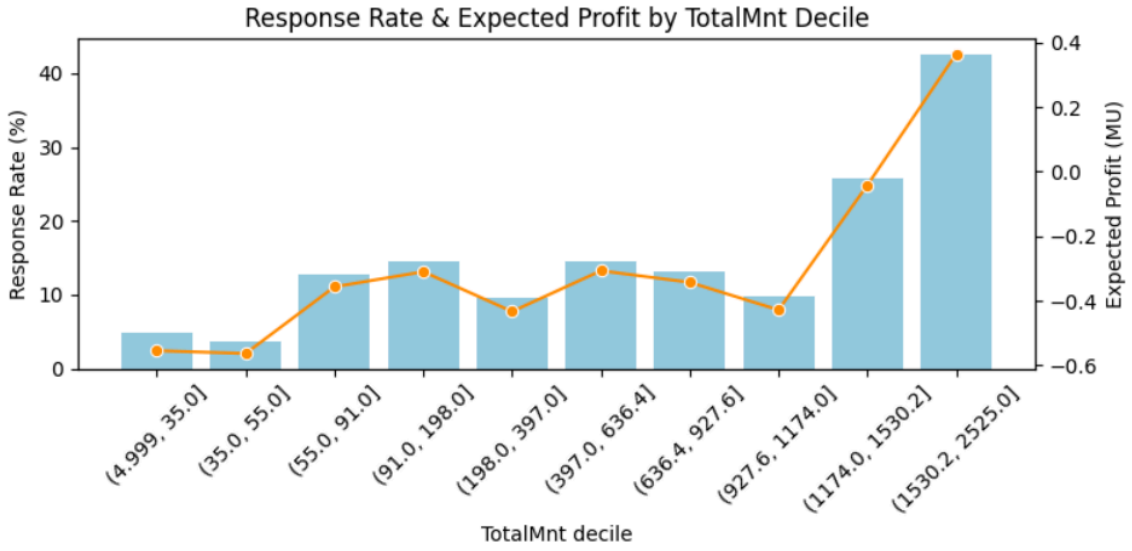
Education and **Marital status** emerged as the main drivers of response. Response rates climb steadily from around 4% for Basic education to over 20% for PhD holders, suggesting I should highlight higher-education levels. Among marital segments, Singles and Widows demonstrate roughly double the engagement (~23 – 24%) of the Married/Together majority (~10 – 12%). By contrast, the country offers little differentiation outside of the predominant Spanish base—most major markets cluster around the 15% average. In countries with small customer bases, like Mexico, spikes are likely statistical noise. These findings will shape the further feature engineering for predictive models: one-hot encode education and marital status to capture their lift, while collapsing or dropping low-volume country categories. Together, these plots pinpoint the key individual drivers of campaign success, laying the groundwork for our subsequent binned response 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, I 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 I account for both per-contact costs and revenue from conversions.



The Recency decile analysis reveals a striking non‑linear relationship between days since last purchase and both response likelihood and profitability. In the lowest decile (0 to 71 days), very recent buyers respond at only around 2%, producing a hefty net loss (≈ -0.48 MU) when contacted. Moving into the middle deciles (140 to 424 days), response rates rise modestly (8% to 14%) and losses shrink (≈ -0.45 MU to -0.32 MU), but these segments remain unprofitable. It is only in the top two deciles—customers with Recency beyond roughly 560 days—that response surges to 24–32% and expected profit turns positive (≈ +0.02 MU). This clear "threshold effect" shows that targeting the most dormant customers is the only profitable strategy when considering campaign costs, directly guiding our model to prioritize long-inactive but high-potential buyers above all others.

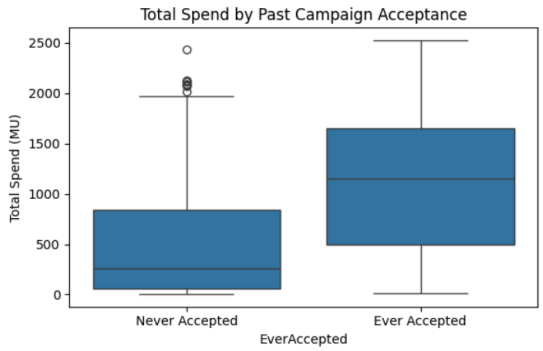


The TotalMnt decile analysis uncovers a similarly pronounced non‑linear effect of historical spending on both response likelihood and profitability. In the lowest spend deciles (0 – 55 MU), customers respond at under 5% and contacting them yields a steep loss (≈ -0.55 MU). As spending rises into the middle tiers (55 – 400 MU), response rates improve to 10 – 14%, but expected profit remains negative or barely breaks even (≈ - 0.4 to 0.3 MU). Even higher spenders (400 – 1,174 MU) maintain response around 12 – 14% with only incremental profit gains, still below true profitability. It is only among the top two deciles (spending >1,174 MU) that response jumps dramatically, reaching ~26% in the 9th decile and ~43% in the top decile and driving strongly positive profits (+0.1 to +0.4 MU). This clear cutoff confirms that only the

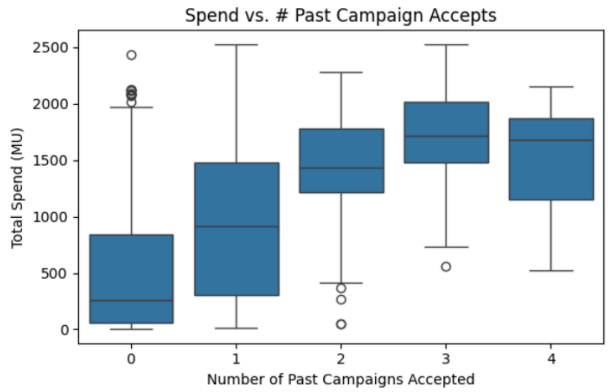
highest‑spending customers generate a net gain and should be the primary focus of marketing outreach.

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. This supports our motivation for predictive models that aim to find a small portion of the targeted customer base to maximize the effectiveness of campaigns.

3.3 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
```

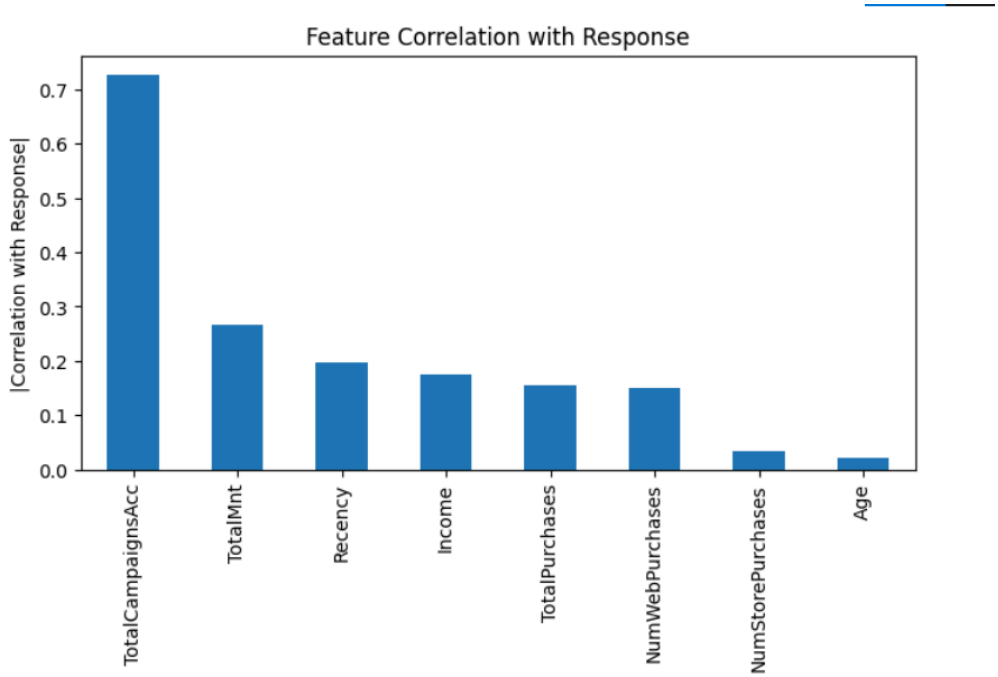



```
Correlation between # past accepts and total spend: 0.458
```

Building on our RFM findings, I examined past campaign engagement as a loyalty and value signal. I created two features—**EverAccepted** (any past campaign response) and **NumPastAccepts** (0–5 total accepts)—and compared them against total historical spend. Customers who had ever responded spent on average 2 – 4× more (mean ~1,092 MU vs. 480 MU) than those who never did, and spending rose steadily with each additional past acceptance

(median jumps from <300 MU at 0 accepts to ~1,300 – 1,700 MU for 2 – 4 accepts). A correlation of 0.458 between NumPastAccepts and TotalMnt confirms this strong link. These patterns reinforce that past responsiveness predicts future response and marks high‑value customers, making **TotalPastAccepts** an essential feature in our predictive model.

3.4 Correlation with Response & Redundancy Check



Building on our RFM and loyalty insights, I quantified each numeric feature's direct association with response by ranking their absolute correlations. **TotalCampaignsAcc** led the pack, reaffirming past accepts as the strongest loyalty signal. **TotalMnt** and **TotalPurchases** followed, highlighting monetary and frequency power. **Recency** ranked third, echoing our finding that dormant customers respond best. Other features—age, Income, and channel‑specific counts—showed modest correlations and can be treated as secondary signals. This again confirms our focus on RFM and engagement metrics for the predictive model.

EDA Conclusion & Transition to Modeling

Our EDA showed that a handful of features—historical spending (TotalMnt), purchase frequency (TotalPurchases), dormancy (Recency), and past campaign engagement

(EverAccepted/NumPastAccepts)—are by far the strongest discriminators of responders versus non-responders, with supplemental lift from higher education and single/widow status. These insights directly inform our modeling: I will build on RFM and loyalty metrics as core predictors, incorporate Income and Education flags to refine the forecast, and exclude or consolidate weaker, redundant signals (e.g. raw Age or low-volume country codes). By translating the clear thresholds and non-linear "knots" uncovered in EDA into engineered features, ensure our predictive models focus on the customer behaviors that truly drive campaign profitability.

4. Predictive Models and Methods

To translate our exploratory insights into a practical targeting strategy, I 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: 1. data preparation, 2. preprocessing, 3. model training, hyperparameter tuning, and evaluation, 4. model application

1. Data Preparation

I 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, I excluded identifiers (ID, Dt_Customer) and indicators directly related to past campaigns (AcceptedCmp1 to AcceptedCmp5). I performed an 80/20 train-test split of the dataset into training and testing subsets, preserving the original response rate distribution (~15%) in both subsets, ensuring reliable generalization to unseen data.

2. Preprocessing Pipeline

To ensure our models to handle mixed data types uniformly, I built a ColumnTransformer that wrapped two pipelines. For numeric predictors, I first applied a median imputer to fill any missing values and then standardized each feature to zero mean and unit variance. For categorical fields (Education, Marital_Status, and Country), I imputed missing categories with

the most frequent level and then one-hot encoded the values, dropping one level for binary variables and ignoring any unseen categories.

3. Model Training, Hyperparameter Tuning, Model Evaluation

The analysis evaluated five classifiers to find the best-performing models: **Logistic Regression**, **k-Nearest Neighbors**, **Decision Trees**, and **XGBoost**. To guard against overfitting and ensure robust performance, I adopted a two-stage validation strategy. First, I conducted 5-fold stratified cross-validation with GridSearchCV on the training set, optimizing ROC-AUC to rank customers by response likelihood under class imbalance. After identifying each model's optimal hyperparameters, I retrained its full pipeline on all training data and then performed an out-of-sample evaluation on our held-out validation set, measuring ROC-AUC, precision, recall, and F1-score. This approach gave us both the best model choice (XGBoost) and a clear, unbiased view of each algorithm's real-world effectiveness.

4. Further Evaluation & Application

Beyond ROC-AUC, I considered the profit-maximizing threshold, translating predicted probabilities into actionable decisions. I 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 the accuracy scores.

5. Results and Interpretation

5.1 Model Selection and Validation Performance

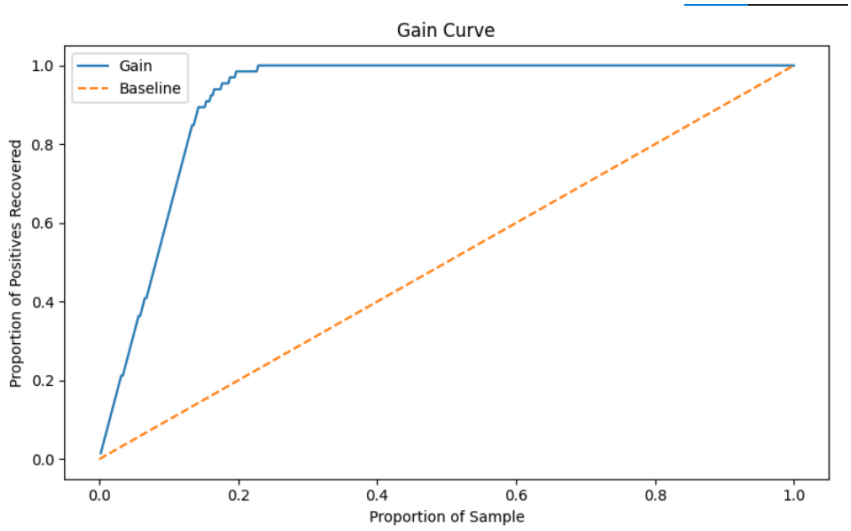
Validation Performance of All Models:

	model	ROC-AUC	Precision	Recall	F1-score
4	xgb	0.990505	0.893939	0.893939	0.893939
0	logistic	0.989455	0.920000	0.696970	0.793103
3	rf	0.987374	0.913793	0.803030	0.854839
2	tree	0.983657	0.846154	0.833333	0.839695
1	knn	0.953576	1.000000	0.560606	0.718447

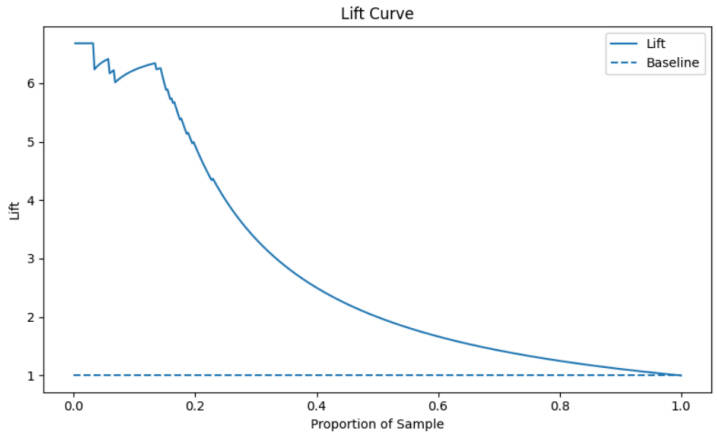
After rigorous evaluation using a robust validation methodology, I selected **XGBoost** as our final predictive model due to its superior predictive performance on unseen validation data. It achieves the highest ROC-AUC (0.9905) and balancing precision (89.4%) and recall (89.4%) for an F1 of 0.893. While Logistic Regression closely trailed in AUC (0.9895) and boasted slightly higher precision (92.0%), its lower recall (69.7%) limited its overall effectiveness. The Decision Tree model reversed that trade-off, strong recall (83.3%) at lower precision (84.6%), and Random Forest performed well but fell short of XGBoost's ranking power. Last, KNN lagged behind on AUC (0.9536) despite perfect precision. These results confirm XGBoost as the best choice for accurately ranking likely responders and driving profit-optimized targeting.

5.2 Gain and Lift Curve Analysis

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



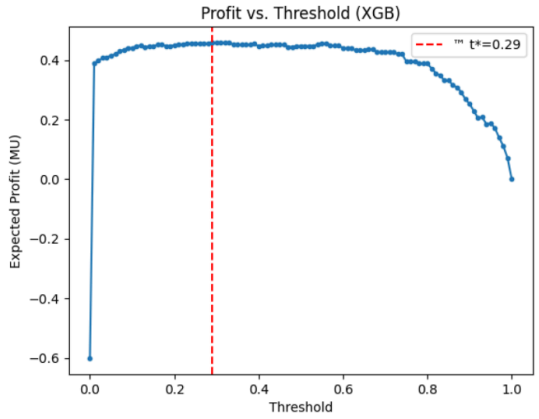
Gain Curve: The XGBoost gain curve shows that nearly **100%** of responders lie in the **top 20%** of scored customers, delivering a more than five-fold improvement over random targeting. In practice, this means iFood can cut outreach to one-fifth of the list and still capture virtually every sale, drastically reducing contact costs and multiplying return on marketing spend.



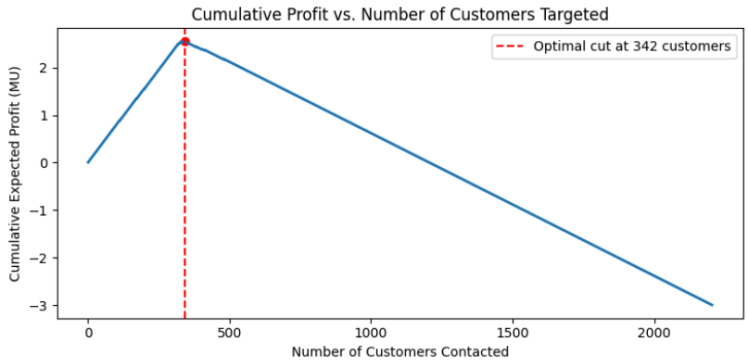
Lift Curve: The lift curve peaks at **6×** in the top decile, meaning those top-scoring customers are six times more likely to respond than random picks. As the lift decreases deeper into the list, it underscores that precision targeting of the highest-ranked segment yields the greatest success, far outpacing mass-marketing.

5.3 Expected Profit-Driven Threshold Optimization & Final Customer Selection

```
Optimal threshold = 0.29, Expected profit = 0.5 MU
```




```
Customers to target next month: 346
```



Despite strong ranking metrics, our true objective was to maximize campaign profit. To bridge predictive scores and business outcomes, I first performed a profit threshold on our validation set, incorporating real outreach costs and per-responder revenue. Then, found that a probability cutoff of **0.30** would, in theory, target **346 customers** and yield **0.5 MU** in expected profit. With this threshold validated out-of-sample, I then applied it to the full customer base, ranking all 2,205 customers by predicted response probability and computing cumulative profit. This finer-grained analysis revealed that contacting the **top 340** prospects (≈15.4% of the base) maximizes total expected profit at **2.54 MU**, turning a prior 3.046 MU pilot loss into a substantial gain. Together, iFood can use this optimal customer base as the target for the next campaigns to maximize campaign effectiveness and profits.

```
Optimal number of customers to contact: 342
Which is 15.5% of the entire base
Maximum cumulative profit: 2.55 MU
```

	ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines	...	Year_Customer	TotalMnt	TotalPurchases	TotalCampaignsAcc
0	10446	1957	PhD	Married	82017.0	0	0	2012-11-07	58	184	...	2012	729	17	4
1	6906	1953	Master	Widow	84953.0	0	0	2013-06-03	73	167	...	2013	1024	18	4
2	7872	1975	PhD	Married	86836.0	0	0	2012-09-12	7	179	...	2012	557	23	4
3	8443	1972	Graduation	Single	24762.0	1	0	2014-02-10	16	6	...	2014	86	9	2
4	4692	1976	Graduation	Married	7500.0	1	0	2012-08-01	19	7	...	2012	71	12	2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
337	5350	1991	Master	Single	90638.0	0	0	2014-02-13	29	1156	...	2014	2525	18	2
338	10675	1956	PhD	Married	66334.0	0	1	2013-04-03	82	909	...	2013	1161	19	1
339	8975	1968	Graduation	Married	19514.0	1	1	2014-01-26	47	14	...	2014	69	10	1
340	4637	1954	PhD	Single	74637.0	0	0	2013-05-18	73	960	...	2013	1650	25	1
341	4310	1944	Graduation	Married	80589.0	0	0	2014-01-22	25	507	...	2014	1428	21	2

342 rows × 39 columns

6. Conclusion & Recommendations

This project showed that a targeted, data-driven approach can turn a loss-making pilot into a profitable campaign. Our EDA identified “lapsed but loyal” customers – those with high historical spend, frequent purchases, long recency, and past campaign engagement – as the most responsive. After comparing five classifiers, XGBoost emerged as the best (ROC-AUC≈0.99), and our gain, lift, and profit-threshold analyses revealed that contacting just **340 customers** (≈15% of the base) would flip a –3.046 MU pilot loss into an estimated **+2.54 MU** profit. These elite targets – predominantly single, well-educated adults with strong spending histories – form the ideal focus for the next gadget offer.

One further validation with the real data could be an A/B test against a random or legacy control group to confirm real-world uplift and profit. Feed the outcomes back into the model for continuous retraining to improve our optimal customer base and convert to loyal customers by precision-targeted campaigns, maximizing long-term profitability. Besides, a key limitation of this project is that it considers only the existing customer base: our model cannot identify potential new customers or address acquisition strategy. Future research should therefore extend beyond retention and reactivation to new-prospect targeting to acquire new customer bases. This would complement our current work and help iFood grow its overall customer funnel, not just optimize existing contacts.