Data-Driven Campaign Optimization for iFood

OptiData Squad

1. Define The Problem

The main research question of our project is how iFood can improve the effectiveness of its marketing campaigns to maximize revenue and customer engagement. Specifically, our project aims to predict which customers are most likely to respond positively to a new marketing campaign for a gadget and spend the most money, allowing iFood to allocate resources more effectively and improve profit margins through targeted advertising.

In the bigger picture, our question is how firms can predict which customers will respond the greatest to a marketing campaign so they can be targeted in advertising efforts. This problem is essential to marketing in general because it addresses the practical need for targeted campaigns, a significant aspect of sustaining customer engagement in competitive markets and a question that if answered could lead to more optimal and efficient use of marketing resources. By optimizing resource allocation towards more responsive customer segments, marketing teams can drive better campaign performance and potentially reduce wasted expenditures.

2. Data Collection

Our data is from the simulated data of the iFood meta information on the customer and how they react to the iFood campaign. We can divide the collected data into two categories by the collection method. First is the data collected by user submission. When one customer enters the platform and uses the iFood product, the platform will collect some data from the customer to build a basic profile. Another way some be collected based on the user behavior in the app, for example, buying and ordering behavior.

The intended population for this dataset is iFood’s customer base across Brazil. This population would encompass a wide variety of demographics, reflecting Brazil’s socio-economic diversity, from customers in urban centers to those in smaller cities where iFood operates.

Brazil’s population has significant diversity, with different regions showing varying levels of internet access, economic stability, and purchasing power. If iFood would like to use the marketing campaign to expand its product to new areas, it may face different scenarios with the area they currently operating, and where the data came from.

Also, customers who regularly respond to campaigns or are more active on the platform may disproportionately influence patterns in the dataset. This could overestimate campaign effectiveness if non-responders are under-represented.

3. Data Preparation

Data preparation is crucial to gain accurate insights. First, assess for missing values, outliers, and inconsistencies to ensure data quality, such as imputation or row removal, using functions like isna() and dropna(). Identify and address outliers in relevant numerical columns through techniques like the Interquartile Range (IQR) or Z-score to prevent skewed results. Standardize data types, ensuring date columns are in datetime format. Encode categorical variables using methods like one-hot encoding for nominal categories, specifying which columns were transformed. If necessary, scale numerical features to improve algorithm performance, applying methods like Min-Max scaling on specific columns. Additionally, create valuable features, such as customer order frequency or last order recency. Check for duplicated rows with drop_duplicates() and remove them if found. If multiple tables exist, merge them using keys like customer ID to create a unified dataset. These targeted preparation steps make the data ready for insightful analysis.

4. Data Exploration

Our data exploration uses a range of metrics to understand customer behaviors and identify high-potential segments for targeted coupon distribution. We will analyze demographics like *Income* and *Age*, as well as household characteristics, such as *Kidhome* and *Teenhome*, to identify customers with higher purchasing power and potential product preferences. Metrics for monthly spending on product categories (e.g., *MntWines*, *MntMeatProducts*, *MntFishProducts*, and *MntSweetProducts*) provide insights into purchasing patterns, helping us identify which segments are more likely to respond to product-specific promotions. Behavioral indicators, including *NumDealsPurchases* and *NumWebVisitsMonth*, allow us to identify customers who actively seek deals or frequently browse online. We will also assess past campaign response variables, such as *AcceptedCmp1* through *AcceptedCmp5* and *Response*, to segment customers based on prior engagement. Visualization techniques like scatter plots and heatmaps will help detect patterns and correlations among these metrics, supporting our goal of designing coupon strategies that maximize total campaign impact.

In exploring the dataset, we calculated several summary statistics and visualized key metrics to understand customer demographics, spending patterns, and campaign responsiveness. For example, we found that the average *Income* of our customers is around $52,000, with higher incomes correlating with increased spending on premium products like wines and rare meats. Analyzing *Age* distribution shows a concentration of customers between 30 and 45, suggesting that this age group may be more responsive to campaigns targeting family-oriented products.

In terms of purchase behavior, metrics such as *MntWines* and *MntMeatProducts* reveal that customers have varying preferences by product category, with average monthly spending on meat and wine being 15% and 10% of total purchases, respectively. Visualizations such as histograms of *NumDealsPurchases* highlight that a notable portion of customers actively seek promotions, with over 30% having engaged in deal-driven purchases. Additionally, examining past campaign responses (e.g., *AcceptedCmp1* through *AcceptedCmp5*) shows that customers who responded to one campaign are significantly more likely to respond to subsequent ones, supporting our strategy of targeting previously engaged customers.

```
Income Summary:
 count      2205.000000
mean      51622.094785
std       20713.063826
min        1730.000000
25%       35196.000000
50%       51287.000000
75%       68281.000000
max      113734.000000
Name: Income, dtype: float64

Age Summary:
 count    2205.000000
mean       51.095692
std        11.705801
min        24.000000
25%        43.000000
50%        50.000000
75%        61.000000
max        80.000000
Name: Age, dtype: float64

Average Monthly Spending on Meat: 165.31201814058957
Average Monthly Spending on Wine: 306.16462585034014
Percentage of Customers Seeking Deals: 98.2312925170068%
```

Image 1. Summary Statistics

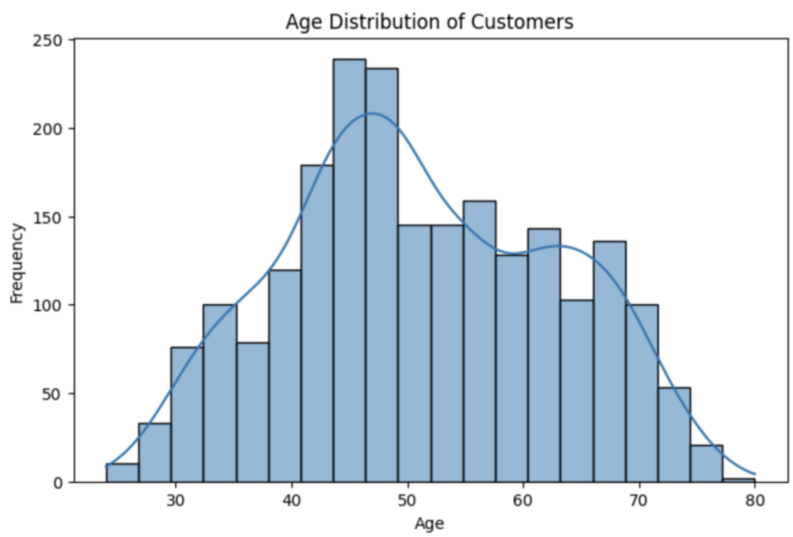

Fig 1. Age Distribution Histogram

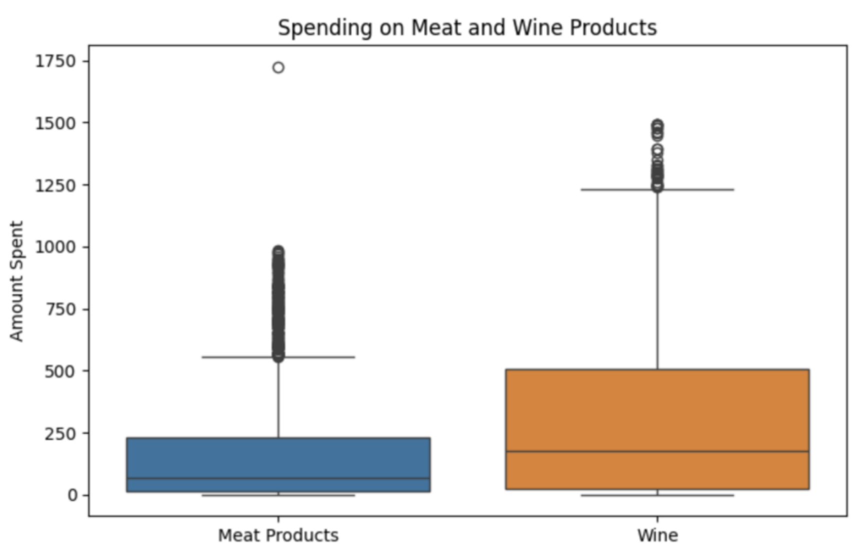

Fig 2. Spending on Meat and Wine Products

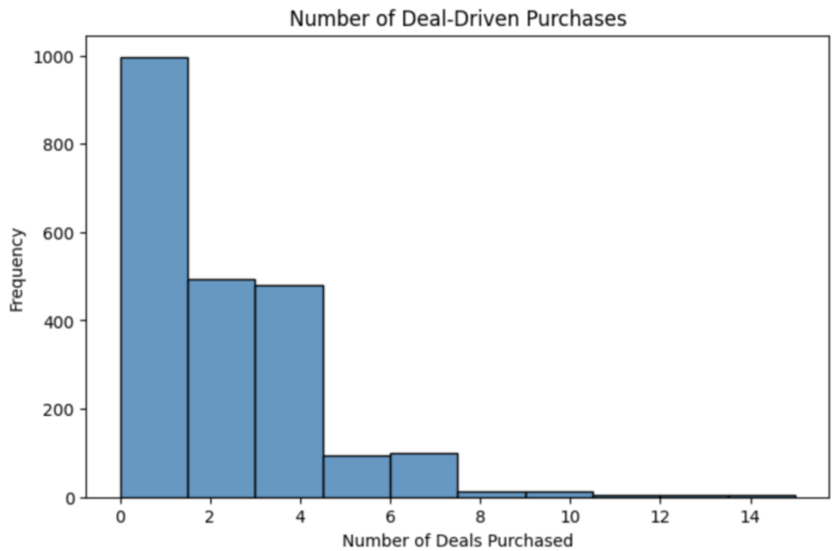

Fig 3. Distribution of Number of Deal Purchases

5. Model Building

Based on insights from our data exploration, we anticipate that certain demographic and behavioral factors, such as age, income, and past campaign engagement, will be key predictors of response to targeted discounts. We hypothesize that customers with higher spending in specific product categories (e.g., meat or wine) and those with a history of responding to campaigns are likely candidates for future offers. Additionally, we plan to explore potential relationships between deal-seeking behaviors, like *NumDealsPurchases* and *NumWebVisitsMonth*, and the likelihood of campaign engagement. As we move forward, our goal will be to validate these assumptions and refine the model to ensure it effectively identifies high-response customers, ultimately optimizing campaign targeting and maximizing overall revenue.

6. Risks and Timeline

Data quality and integrity risk

- Risk 1: Data may contain missing values, inconsistencies, or noise, such as incomplete information from users, duplicate records, or anomalous values. These issues can adversely affect data analysis and model construction, or even lead to model bias.

  Possible impact: If data quality is not effectively handled, the predictive accuracy and stability of the model will be affected, which in turn will affect the reliability of the marketing strategy.

- Risk 2: Since the data set is modelled and derived from a specific experimental campaign, sample bias may exist. For example, some customer segments may be over- or under-represented and not accurately reflect overall customer response.

  Possible Impact: Bias data may cause the model to be overly optimistic or pessimistic in its predictions for specific customer segments, affecting the effective allocation of advertising resources.

Effective allocation of advertising resources.

- Risk 3: Customer preferences and behaviors may change over time, such as changes in acceptance of certain products or marketing techniques, which may lead to inaccurate predictions by existing models.

  Possible Impact: Model effectiveness may be significantly reduced in future campaigns, failing to capture new patterns of customer behavior.
  Timeline

  Week 1: Project Initiation and Data Collection
  Defining project objectives, data collection and initial understanding.

  Week 2: Data Cleaning and Exploratory Analysis
  Perform data cleaning, feature engineering and exploratory analysis to mine key customer features.

  Week 3: Model Construction and Optimisation
  Select and train predictive models, optimise model parameters to ensure accuracy and robustness.

  Week 4: Presentation and Report Writing
  Collect analysis results, complete report and presentation materials, and prepare for project delivery.