# Machine Learning Team Assignment 4

Group 12 Members: Jenny Chiu, Conor Qiao, Emaleigh Neo

The analysis examines the performance of a marketing campaign conducted by a Brazilian retail food company, which targeted 2,240 randomly selected customers across various product categories, including wine, rare meat, exotic fruits, specialty fish, and sweets. These products are divided into two tiers: gold (premium) and regular. Sales were conducted through three distinct channels: physical stores, catalogs, and the company's website. The campaign involved contacting each customer six times via phone to encourage purchases, with responses recorded for each call. Despite being cost-effective on a per-contact basis, with an average cost of $3 per customer, the campaign ultimately resulted in an overall financial loss. To learn from this initiative, we analyzed the resulting data to extract key features and reduce complexity through principal component analysis. We then segmented customers into clusters based on their purchase behaviors, assessed their responses to discounts, and developed a predictive model to identify high-potential customers for future campaigns, aiming to enhance the marketing effectiveness and profitability.

## Methods

To enhance our understanding of customer behavior, we first mapped the categorical variables to integers and conducted feature engineering. Key features included purchase behavior metrics such as NumWebPurchases and NumDealsPurchases, along with responses to past campaigns. These features were standardized before applying Principal Component Analysis (PCA) for dimensionality reduction. We selected 16 components, as this captured 90% of the cumulative explained variance, ensuring minimal information loss. Subsequently, K-means clustering was performed on the PCA-transformed data, with the optimal number of clusters determined to be 3 through the Elbow method, where the rate of decrease in the graph slowed, complemented by Silhouette scores to ensure significant segmentation.

We analyzed the clusters based on purchase behavior, campaign responses, and discount sensitivity, providing valuable insights into customer preferences and profitability potential. Key profitability metrics, including total profit, profit per customer, and cluster reach, were calculated. Targeting specific clusters was simulated to identify high-value segments for future campaigns. Finally, a regression model was built using XGBoost to predict total purchases per customer. Customers with the highest predicted purchase likelihood were targeted using the decision rule where their predicted campaign purchases were greater than the cost of the campaign per customer as we want a return larger than the cost per customer. The resulting profits from this approach were compared with those from the cluster-based strategy to assess the most effective method.

## Results and Discussion

## Results

This analysis utilized PCA for dimensionality reduction and K-Means clustering to simplify complex data into 16 key features, successfully retaining approximately 90% of the variance. Through this approach, we identified three distinct customer groups with significant differences. Figure 1 illustrates the elbow method, used to select the number of clusters and figure 2 shows the clustering. Additionally, we built a predictive model to predict purchases made in a similar campaign, which achieved an RMSE of 2.33. Figure 3 illustrates the feature importance from the model. Table 1 shows the results of purchasing behavior, Table 2 shows the campaign response, and Table 3 shows discount sensitivity by cluster. Cluster 1 had the most purchases overall and was the most profitable during the campaigns, but Cluster 2 was the most responsive to deals. Table 4 shows the profits per cluster, Table 5 shows profits from customers targeted by the model, while Table 6 directly compares the model with the most profitable cluster. Overall, the customers targeted by the model generated more profit, with fewer people from the customer base invited.

**Table 1: Purchasing behavior by cluster**

| Cluster | Total Purchases | Wine spend | Fruits spend | Meat spend | Fish spend | Sweet spend | Premium spend | Web purchases | Catalog purchases | Store purchases | Preferred channel |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 934,158 | 4,323,685<br>44% | 46,009<br>5% | 295,349<br>31% | 66,427<br>8% | 47,722<br>6% | 54,966<br>6% | 3943 | 4085 | 8.61 | Web |
| 2 | 343,056 | 233,819<br>65% | 7,055<br>2% | 57,166<br>18% | 9,069<br>3% | 6,812<br>2% | 29,135<br>0.10 | 3651 | 1465 | 5.67 | Web |
| 3 | 68,065 | 18,579<br>27% | 5,341<br>8% | 17,548<br>8% | 7,909<br>11% | 5,362<br>8% | 13,326<br>0.20 | 1459 | 369 | 3.14 | Web |

**Table 2: Campaign response by cluster**

| Cluster | Total Camp purchases | Camp 1 | Camp 2 | Camp 3 | Camp 4 | Camp 5 | Camp 6 | Response Rate | Camp profit ($) |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 604 | 125 | 19 | 54 | 78 | 151 | 177 | 14% | 71.68 |
| 2 | 257 | 17 | 11 | 51 | 86 | 11 | 81 | 5% | -1414.81 |
| 3 | 133 | 0 | 0 | 58 | 0 | 0 | 75 | 3% | -1656.89 |

**Table 3: Discount sensitivity**

| Cluster | Deal Purchases |
|---|---|
| 1 | 1,182 |
| 2 | 2,626 |
| 3 | 1,341 |

**Table 4: Profit per cluster**

| Cluster | Total profit ($) | Profitability ($) | Reach |
|---|---|---|---|
| 1 | 71.68 | 0.10 | 32% |
| 2 | -1,414.81 | -1.80 | 35% |
| 3 | -1,656.89 | -2.32 | 32% |

**Table 5: Model Targeted**

| Total profit ($) | Profitability ($) | % invited |
|---|---|---|
| 398.06 | 11.06 | 8% |

**Table 6: Comparison between Model and Cluster 1**

| | Total profit ($) | Profitability ($) | % invited/reach |
|---|---|---|---|
| Model | 398.06 | 11.06 | 8% |
| Cluster 1 | 71.68 | 0.10 | 35% |

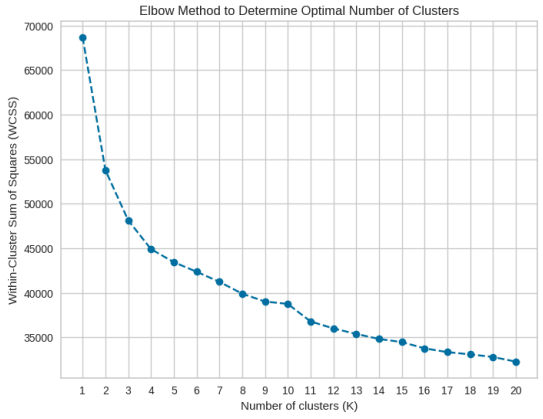


Fig 1:Elbow Method

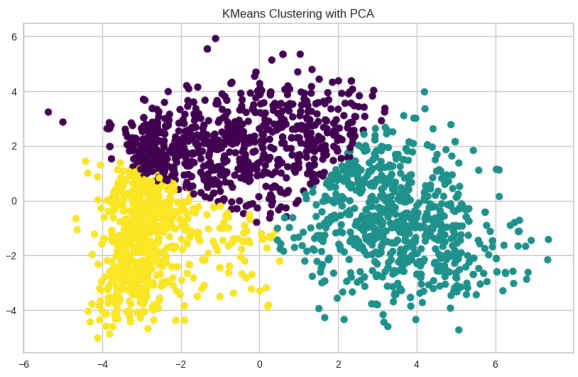


Fig 2: K-means Clustering graph with PCA

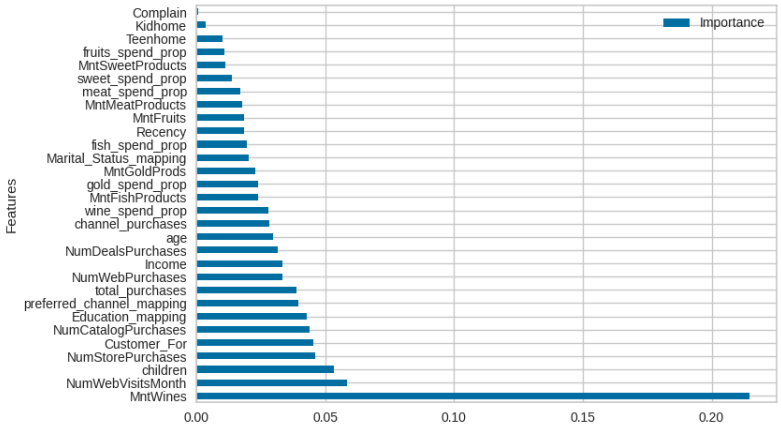


Fig 3: Feature importance graph from XG Boost regression

## Discussion

This project shows how data can help make smarter decisions when planning marketing campaigns. By using techniques like clustering and predictive modeling, we gained valuable understanding about customer behavior and likelihood of their responses to ad campaigns.

Using PCA to reduce the number of features made it easier to focus on the most important patterns in the data. We selected 16 components that together accounted for 90% of the variance, enabling us to focus on the most significant patterns while simplifying the analysis. The clustering results highlighted distinct differences among customer segments, confirming that responses to marketing efforts vary significantly across groups. The optimal number of clusters were determined through the Elbow method graph where the WCSS drop slowed and distance became even. Notably, Clusters 1 and 2 exhibited higher purchasing behaviors; however, Cluster 1 showed the highest campaign-driven purchases, suggesting a key target audience for improving return on advertising. Despite their higher purchase levels, Cluster 1 was less responsive to discounts. Additionally, we discovered that wine and meat were the top choices across clusters, with a uniform preference for online purchasing. These insights are pivotal for crafting more personalized and effective marketing campaigns.

We also built a predictive model to estimate potential purchases for individual customers, enabling us to focus on high-value targets. This approach improved precision, prioritizing customers most likely to convert. Among the factors influencing purchase predictions, the quantity of wine purchased emerged as the most significant, followed closely by the frequency of web visits in a month, against the backdrop of campaign profitability. Our decision rule targeted customers with projected campaign purchases exceeding the $3 campaign cost per customer, ensuring a profitable return on advertising spend.

Comparing clustering with predictive targeting demonstrated the strengths of both strategies—clustering provides valuable broad segmentation, while predictive

modeling allows for a more tailored approach that maximizes campaign outcomes. Importantly, this model also enabled us to reduce the audience size for ads, lowering overall costs while enhancing profitability.

## Conclusion

In conclusion, understanding customer behavior and leveraging data are essential for optimizing marketing campaigns. By employing advanced analytics techniques, such as PCA and K-Means clustering, we have successfully identified key customer segments that can drive profitability. Future marketing efforts should focus on allocating resources to high-value groups, like Cluster 1, while minimizing advertisements in less profitable segments, such as Cluster 2 and Cluster 3. Additionally, integrating predictive models to target high-value customers and continuously updating data will enhance campaign adaptability and sustain long-term growth. This data-driven approach not only prioritizes impactful engagement but also aims to maximize returns on advertising.