**ALY6020 Predictive Analytics**

Dr. Thomas Goulding

Module 6

Final Marketing Campaign Project

Jeff Hackmeister

6/27/2025

**INTRODUCTION**

For this project, customer data will be analyzed to help improve the marketing campaign efficiency for a magazine company’s marketing department. The marketing department has provided a dataset containing records for 2,240 customers with both previous purchasing, as well as demographic, information. From the provided dataset, we can see that only 334 of the customers responded positively to a marketing campaign, that is a success rate under 15%. Given the available information about customers, the goal of the project to identify the best potential new customer base and to build a predictive model to determine the next marketing campaign and benchmark real work results.

**DATA CLEANING AND FEATUER ENGINEERING**

While the data provided from the marketing department is very clean, there are several preparation steps that need to be taken before deeper analysis is possible. To begin, all needed python packages are loaded.

A screen shot of a computer program

AI-generated content may be incorrect.

These are standard data science packages that will allow for customer segmentation and modeling work to be done. A quick check on missing values reveals only 24 missing Income values, these can be easily handled along the way by entering the median income value from the dataset for these customers.

A screen shot of a computer

AI-generated content may be incorrect.

A few new variables will be added to the dataset, derived from the original data :

* Age will be calculated from date of birth
* Total\_Spending will be the sum on all spending categories
* Total\_Purchases will be the sum of all orders places across channels
* Avg\_Spending\_Per\_Purchase will come from total spending and total purchases
* Total\_Children will simply add the kids and teens at home values
* Customer\_Tenure is calculated from the date of customer acquisition
* Total\_Campaigns\_Accepted is the sum of all accepted offers across previous campaigns
* Preferred\_Channel is the channel with the greatest number of orders
* Premium\_Spending is the total amount spent on wine and gold products
* Regual\_Spending is the total amount spent on the remaining categories

The next engineering step is to encode (or add numerical values) for the categorical values for education and marital status, this will ensure these values are included in the analysis. The final preparation step is to scale the values, centering them all around a mean of zero, to ensure each variable is given equal weight in the analysis, rather than variables naturally containing larger numbers (such as Total Spending) having more influence that variables with lower numbers (such as Total Children).

**CUSTOMER CLUSTERING**

Cluster analysis is a classic methodology in marketing analysis. By better understanding different segments of a customer base, firms are better able to customize campaigns to each cluster. These are often much more successful and efficient marketing investments, especially as compared to general population campaigns. To determine the optimum k value, two different methods will be used – elbow and silhouette. The elbow method calculates the within-cluster sum of squares (WCSS), also known as inertia, for cluster solutions ranging from k=2 to k=10. When visualized, the optimal k is identified at the "elbow" point where the rate of inertia decreases begins to level off. For silhouette, scores for each clustering solution are measured on how similar objects are to their own cluster compared to others. Silhouette scores range from -1 to +1, where higher values indicate better-defined clusters. The optimal number of clusters corresponds to the k value that maximizes the average silhouette score. The results of each are seen here.

A graph of a graph of a graph

AI-generated content may be incorrect.

After testing cluster solutions up to k=10, k=2 was chosen. Both elbow method and silhouette analysis indicated k=2 as optimal. A two-cluster solution aligns with common marketing frameworks (high-value vs. budget-conscious customers). Additionally, this binary segmentation provides clear, implementable marketing strategies 4.

With the K value optimized, the K means model can be utilized.

A math equations and formulas

AI-generated content may be incorrect.

Which results in the following clusters

A white screen with black text

AI-generated content may be incorrect.

And this is the division between the two clusters by the key variables.

A screenshot of a computer screen

AI-generated content may be incorrect.

A screenshot of a computer screen

AI-generated content may be incorrect.

A screenshot of a computer screen

AI-generated content may be incorrect.

These splits indicate a classic marketing segmentation between premium and budget customers. The unequal cluster sizes reflect realistic market segmentation, where premium and budget customer segments rarely occur in equal proportions. The 60-40 split suggests a market structure with a substantial mainstream segment and a significant premium segment, both large enough for targeted marketing strategies. See Image 1 in the image appendix for additional information on the clustering breakdown.

**MODELING CUSTOMER BEHAVIOR**

Understanding the breakdown between premium and basic customers among those included in the dataset is vital to maximizing marketing spend but being able to model and predict future customer behavior provides even more value to the organization. The modeling options are many, which calls for a systematic evaluation of multiple modeling techniques. For the evaluation a logistic regression, random forest and gradient boost model will be fit to the data. The results of the all three as a shown.

A close-up of numbers

AI-generated content may be incorrect.

The three models were systematically evaluated using 5-fold cross-validation and holdout testing. Gradient Boosting emerged as the optimal model with test Area Under the Curve (AUC) of 0.863, closely followed by Random Forest (0.851) and Logistic Regression (0.784). AUC measures the model's ability to distinguish between campaign responders and non-responders across all classification thresholds. An AUC of 0.863 indicates that the model correctly ranks a randomly selected responder higher than a randomly selected non-responder 86.3% of the time, demonstrating excellent discriminatory performance for marketing applications

The ensemble methods demonstrated substantial improvements over the linear baseline (AUC = 0.08), indicating that non-linear customer behavior patterns require sophisticated modeling approaches. Knowing that the gradient boost performed the best, additional diagnostic statistics were evaluated.

A screenshot of a computer program

AI-generated content may be incorrect.

This shows the disparity in performance between the model’s ability to identify customers who do not respond to an offer (precision 0.88 and recall 0.98) versus customers who do respond (precision 0.64 and recall (0.21). This is a highly conservative model but should be very helpful in maximizing the cost per acquisition of future marketing campaigns. Even with these conservative predictions, a 64% precision means campaigns targeting model-identified responders achieve 4.3x the baseline response rate (64% vs 15% overall rate).

With the selection of the ensemble modeling method of gradient boost, the dataset will be divided into test and training sets. Then, the model is fit to the training data and the results are used to make predictions on the remaining test data. From these results, it is determined that these are the most impactful variables in the model.

A screenshot of a computer screen

AI-generated content may be incorrect.

A graph of a graph showing a number of red and white bars

AI-generated content may be incorrect.

At the top of the list, recency is the most important variable in the model. Those who have responded to a campaign recently are likely to respond once again. The key indicators of a premium cluster member are also high on the list with Total Spending and Income also in the top 5. In addition to fitting into the previously discussed cluster, these results also present as a classic Recency, Frequency, Monetary (RFM) segmentation model. RFM scoring divides customers into quintiles (5 equal groups) for each dimension: Recency (recent buyers score 5), Tenure (long-term customers score 5), and Monetary value (high spenders score 5). These scores are summed to create a composite score ranging from 3 (lowest value) to 15 (highest value). Applying those ideas to our data, the customer data can be segmented by RFM score.

A screenshot of a computer screen

AI-generated content may be incorrect.

This demonstrates the importance of the top tier RFM customers (RFM >= 12). These customers respond at rates between 27% (RFM 12) to 52% (RFM 15). As the RFM scores decrease, the likelihood of response drops to 0 at RFM less than 3. Understanding these segmentations is critical to making targeted marketing decisions. See images 2, 3, and 4 in the image appendix for more.

**ANALYSIS AND RECOMMENDATIONS**

Based on the results of the modeling and segmentation analysis conducted, a high level of confidence can be had in the recommendations made to the marketing department. The results of the RFM analysis provides a detailed look at the characteristics of the customer base.

A group of graphs with different colored bars

AI-generated content may be incorrect.

In the first graph, the distribution of RFM scores follows a very normal distribution curve while the campaign response rate show exponential growth in the second graph.

A group of graphs with different colored bars

AI-generated content may be incorrect.Next, the pie chart shows the breakdown of customer types based on RFM score. The Champion category (RFM >= 13) accounds for less than 9% of the customer base, but represents the greatest opporunity for the marketing team to gain customers at very low cost. The Loyal (RFM 10-12) and Potential (RFM 7-9) are nearly 74% of the customer base and represent a huge potential if then can be activated by the right offer and the right campaing.

A group of graphs with different colored bars

AI-generated content may be incorrect.

Finally, this chart demonstrates the power of the targeted strategy behind this analysis. Starting at a response rate below 15% in the initial data, the response rate for the top 20% RFM customers exceeds 35% while Champions are nearly 44%.

The integration of customer segmentation and predictive modeling presented in this analysis offers immediate opportunities for the marketing department. Statistical modeling led to a highly useful model, one that obtained 86.3% AUC predictive accuracy and clear customer tiers ranging from 0% to 45% response rates. The marketing team can confidently allocate resources to maximize return on investment (ROI) based on these results. The binary segmentation (premium vs. budget-conscious customers) simplifies strategic planning while the RFM scoring system provides granular targeting precision. Implementing this framework can expect substantial efficiency gains: targeting just 8.7% of customers (RFM 13-15) generates response rates three times higher than baseline while accessing customers with 10x higher lifetime value. As marketing budgets face increasing scrutiny, such evidence-based targeting methodologies become crucial for demonstrating measurable business impact and sustaining competitive market position.

**REFERENCES**

[1] Box, G. E. P., & Norman Richard Draper. (1987). *Empirical model-building and response surface* (p. 424). Wiley.

[2] GeeksforGeeks. (2017, May 2). *K means Clustering – Introduction*. GeeksforGeeks. https://www.geeksforgeeks.org/machine-learning/k-means-clustering-introduction/

[3] GeeksforGeeks. (2024, February 22). *Random Forest Algorithm in Machine Learning*. GeeksforGeeks. https://www.geeksforgeeks.org/machine-learning/random-forest-algorithm-in-machine-learning/

[4] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning : Data Mining, Inference, and Prediction, Second Edition*. Springer New York.

[5] OpenAI. (2024). ChatGPT (Claude 4) [Large language model]. https://claude.ai

‌[6] Starmer, J. (n.d.). *Gradient Boost*[Review of *Gradient Boost*]. YouTube. https://www.youtube.com/watch?v=3CC4N4z3GJc

‌[7] Steele, B., Chandler, J., & Reddy, S. (2016). *Algorithms for Data Science*. Springer.

[8] scikit-learn. (2010). *2.3. Clustering — scikit-learn 0.20.3 documentation*. Scikit-Learn.org. https://scikit-learn.org/stable/modules/clustering.html

[9] scikit-learn. (2019). *Selecting the number of clusters with silhouette analysis on KMeans clustering — scikit-learn 0.21.2 documentation*. Scikit-Learn.org. <https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html>

[10] Talha Şahin. (2023, June 9). What is RFM Analysis? - Talha Şahin - Medium. Medium. https://medium.com/@talha002/what-is-rfm-analysis-2b88600a3cd9

‌

‌

**IMAGE APPENDIX**

Image #1

A screenshot of a graph

AI-generated content may be incorrect.

Image 2

A graph with red and white bars

AI-generated content may be incorrect.

Image 3

A graph of a product

AI-generated content may be incorrect.

Image 4

A graph with red bars

AI-generated content may be incorrect.