OptiMarket: Predictive Marketing Campaign Optimization

DATA642 – Marketing Analytics

Summer 2025

By:

Raghad Abuhijleh

Erik Montaña

Project Objective
To develop a predictive model that accurately identifies customers most likely to subscribe to OptiMarket's $199 premium membership program. The model will be used to target future marketing campaigns more effectively, reducing unnecessary outreach and maximizing return on investment, while minimizing the number of missed opportunities (false negatives)

Problem Statement
OptiMarket has historically relied on random customer selection for its marketing campaigns promoting a premium membership offer. This approach leads to high marketing costs and inefficiencies by targeting many uninterested customers, while potentially missing high-value prospects. To optimize the upcoming sixth campaign, OptiMarket seeks a data-driven solution to predict which customers are most likely to accept the offer. The challenge is to build a classification model that improves targeting accuracy, minimizes campaign costs, and ensures that customers who are likely to respond positively are not overlooked.

Executive Summary

OptiMarket, a retail company offering a premium $199 membership program, sought to improve the efficiency and ROI of its marketing efforts by targeting likely buyers through predictive analytics. Our team developed a classification model using customer demographic and behavioral data to predict which customers were most likely to respond positively to the premium membership offer. Three models were evaluated—Logistic Regression, Gradient Boosting, and Random Forest—with Random Forest demonstrating superior performance, particularly in minimizing false negatives. Model diagnostics are provided in Appendix Figures A1-A3. Based on our findings, we recommend deploying the Random Forest model to strategically target customers in the upcoming sixth marketing campaign, ensuring resources are focused on high-potential leads.

Plan

OptiMarket's goal is to increase conversions for its premium membership offering while reducing marketing costs. The company had previously run five campaigns using random targeting but now seeks a data-driven approach to identify the customers most likely to purchase. The objective of this project was to build a predictive classification model that could effectively distinguish between likely and unlikely buyers based on historical campaign data and customer attributes.

The dataset contained 2,240 customer records with variables spanning demographics, purchasing behavior, and prior campaign responses. Key fields included income, age (derived from year of birth), education level, family composition, product purchase amounts, and acceptance of past campaigns. Our goal was to clean, analyze, model, and evaluate this dataset to support a smarter targeting strategy.

Analyze

We began with data cleaning, addressing missing values in the 'Income' field using median imputation and renaming mis-formatted columns. Duplicate records were checked and removed, and new features were engineered, including 'Customer_Age', 'Total_Kids', and 'Total_Spent' to capture relevant consumer traits.

Categorical variables such as 'Education' and 'Marital_Status' were one-hot encoded. We also created an 'AcceptedAnyCampaign' variable based on whether a customer had accepted any prior campaign. Exploratory Data Analysis (EDA) revealed interesting patterns: older customers and those with higher spending were more likely to respond positively (see Appendix Figure A4 for spending and family size by education levels). Across most marital statuses, acceptors have higher average income than non-acceptors (Appendix Figure A5). Category mix is similar for acceptors and non-acceptors—wines and meat dominate for both—suggesting overall spend level, not product preference, is the stronger differentiator (Appendix Figure A6). A correlation heatmap helped verify multicollinearity risks and informed initial model selection.

Given the classification task and project focus on minimizing false negatives, we tested three models: Logistic Regression (baseline), Gradient Boosting (ensemble), and Random Forest (nonlinear ensemble).

Construct

All models were trained using a 70/30 train/test split. The target variable was the binary 'AcceptedAnyCampaign'. All numerical features were standardized using z-score normalization prior to model training. Tree-based models (Random Forest and Gradient Boosting) do not require feature scaling for performance, but scaling was applied consistently across all models for uniformity and interpretability.

Logistic Regression served as a simple benchmark, achieving 80% accuracy and a recall of 33%(Confusion matix and ROC featured in Appendix Figure A1). Gradient Boosting improved recall slightly (44%) with similar accuracy (Confusion matix and ROC featured in Appendix Figure A3). Random Forest outperformed both, delivering 84% accuracy and 45% recall, with a precision of 69% (Confusion matix and ROC featured in Appendix Figure A2). Feature importance rankings showed spending behavior and campaign history as the most predictive factors.

Execute

Based on the model evaluation, we recommend using the Random Forest classifier to guide the sixth marketing campaign (See Appendix Figure A2 for the test-set ROC curve (AUC = 0.86) and confusion matrix supporting this choice). This model's superior ability to minimize false negatives means fewer likely buyers will be overlooked, maximizing conversion potential and ROI.

Key actionable insights include:

- Target customers with higher spending across product categories.
- Prioritize individuals who have responded to past campaigns.
- Give preference to customers in the 40–60 age range and those without young children at home.

Ethical considerations were addressed by reviewing model fairness. We ensured that no sensitive personal identifiers were used in predictions, and that decisions were based solely on commercial and behavioral variables. Further monitoring is advised to ensure ongoing fairness and performance.

In future iterations, the model could be enhanced with additional features like email open rates, click-through behavior, or social media engagement. A/B testing different thresholds, and campaign messaging strategies may also refine targeting further.

Appendix

Figure A1: Logistic Regression- Confusion Matrix & ROC Curve

Test-set confusion matrix with TN=334, FP=19, FN=52, TP=43 (~80% accuracy). The ROC curve reports AUC=0.79, indicating moderate discriminative power. At the default threshold the model favors precision over recall, missing a notable share of true positives useful as a baseline for later ensemble models.

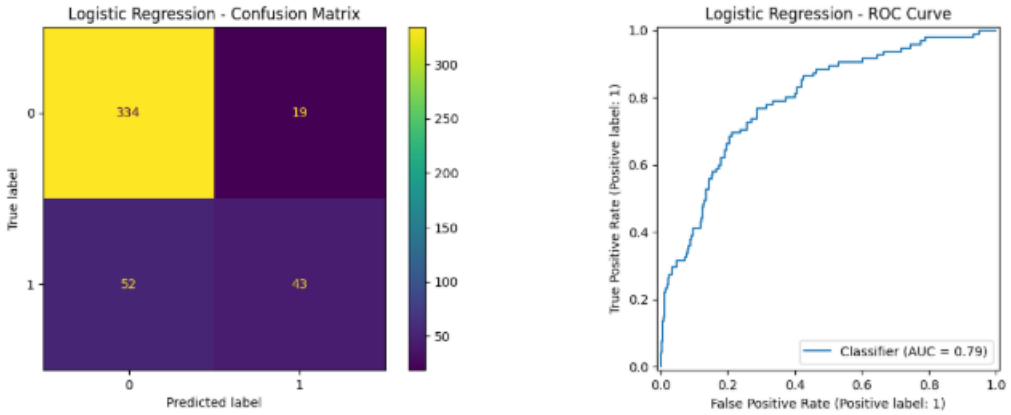


Figure A2: Random Forest - Confusion Matrix & ROC Curve

Test-set confusion matrix with TN=334, FP=19, FN=52, TP=43 (~84% accuracy). The ROC curve shows AUC=0.86, indicating stronger class separation than logistic regression and fewer missed positives for similar overall error supporting selection of Random Forest for deployment.

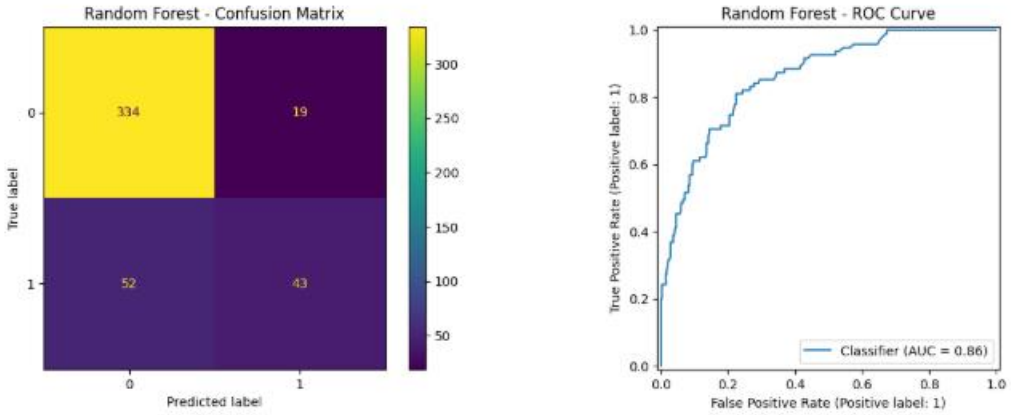
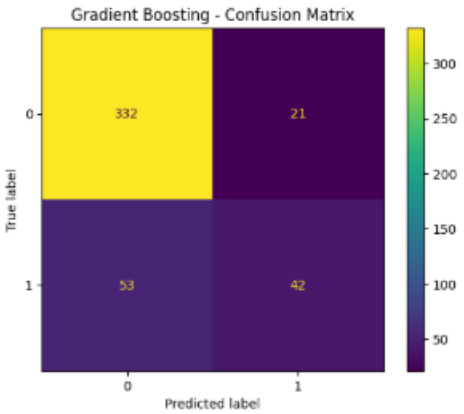


Figure A3: Gradient Boosting- Confusion Matrix & ROC Curve

Test-set confusion matrix with TN=332, FP=21, FN=53, TP=42 (~83% accuracy). The ROC curve shows AUC=0.83 better class separation than logistic regression (AUC 0.79) but below Random Forest (0.86); recall remains limited, so this serves as a solid benchmark model.



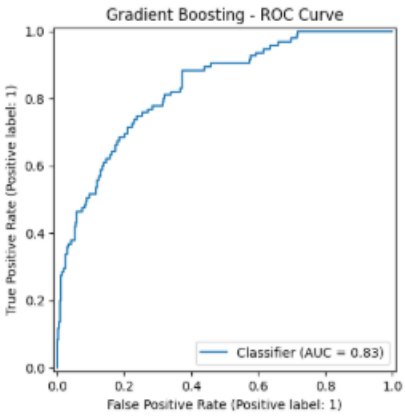


Figure A4: Spending and Family Size by Education Level
Bar chart comparing Total Kids (count) and Total Spending across education tiers. Spending generally rises with education (≈$82 Basic → $497 HS/Some college → $620 Bachelor's → $612 Master's → $672 PhD), while family size peaks among Bachelor's households and is lower at the lowest and mid tiers suggesting higher-education segments tend to spend more.

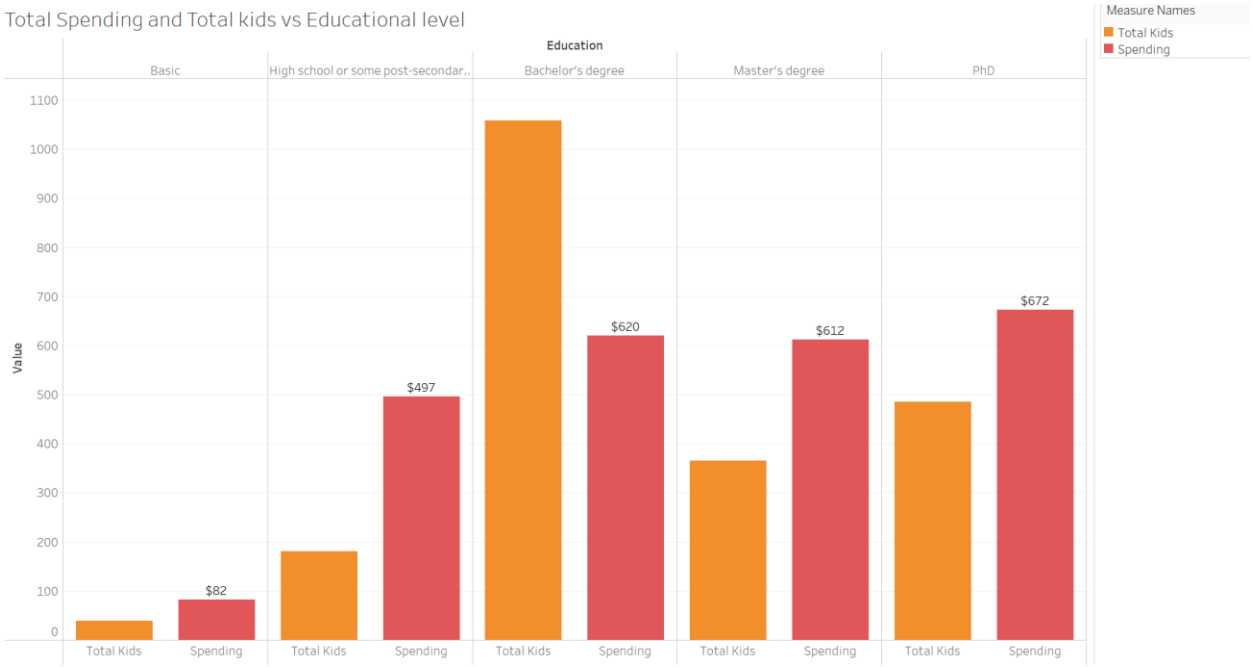


Figure A5: Campaign Acceptance by Marital Status (Average Income Heatmap)
Heatmap shows **average income** by marital status split by whether customers accepted any campaign (darker = higher income). Across most statuses, acceptors have higher average income than non-acceptors (e.g., Married ≈$66k vs $47.9k), indicating a positive association between income and acceptance.

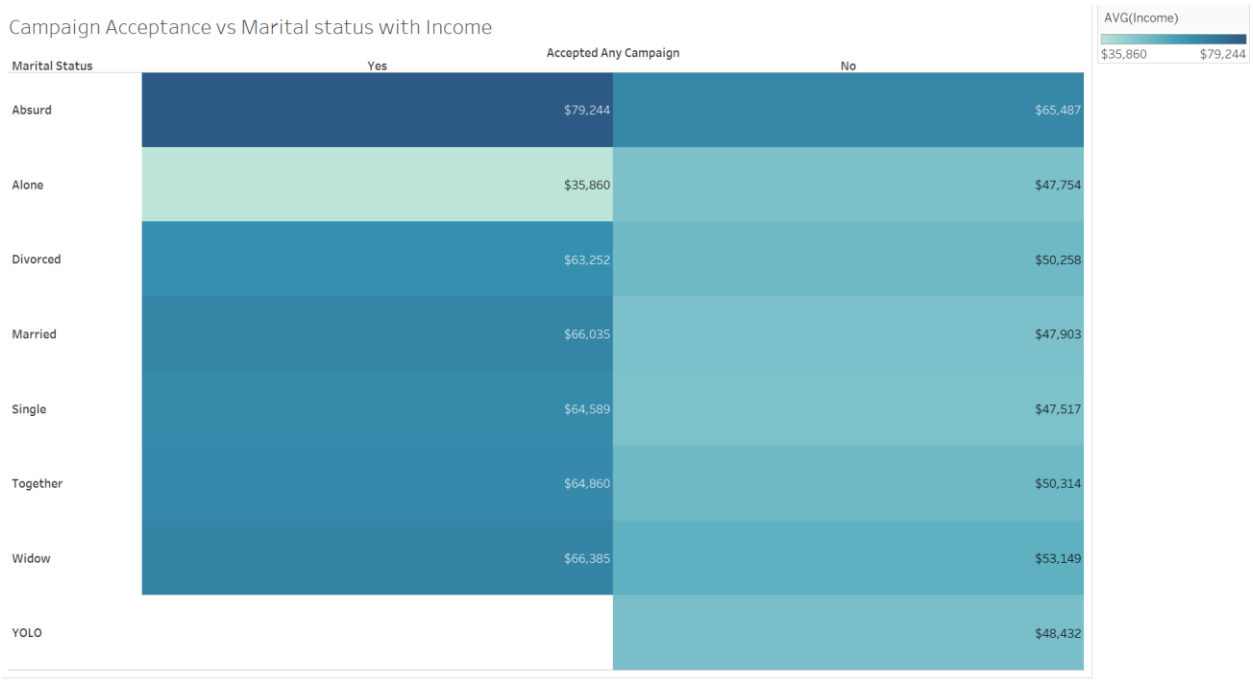


Figure A6. Product Spend by Campaign Acceptance

Grouped bars compare total spend by category for customers who accepted any campaign. **Wines** dominate spend (largest bar), followed by **meat**; other categories contribute far less. Takeaway: category mix is similar across groups overall spend level (not product preference) is the stronger differentiator; normalize by customer count for per-capita insight.