Optimizing Marketing Campaign ROI

Surenther Selvaraj

Bellevue University

DSC550: Data Mining

Professor: Brett Werner

Date: Nov 19, 2025

Term Project Final Write-up Summary: Optimizing Marketing Campaign ROI

Introduction

1. The Problem and Justification

The problem addressed by this project is the **inefficient allocation of marketing budget** in a consumer retail company. The company relies on direct marketing campaigns (e.g., mailers, personalized offers) but suffers from a **low campaign acceptance rate (15%)** and high costs associated with contacting the entire customer base. This broad approach leads to significant wasted spend on customers who are highly unlikely to convert.

It is critical to solve this problem because successful marketing is directly tied to **Return on Investment (ROI)** and overall profitability. By continuing to target 85% of customers who will never respond, the company severely limits its campaign efficiency. A successful predictive model promises to drastically cut campaign volume while maintaining, or even increasing, the absolute number of conversions.

2. Stakeholder Pitch and Data Source

Pitch: "Currently, 85 cents of every marketing dollar is wasted on non-converting customers. Our solution is a **Propensity-to-Purchase Model** that uses behavioral and demographic data to identify the top 10-20% most likely responders. This shift to targeted spending will immediately boost our conversion rate, reduce operational costs, and maximize the return on every dollar we invest in customer acquisition and engagement."

Data Source: The project utilized the **"Customer Personality Analysis"** dataset obtained from Kaggle. This dataset provides a real-world proxy of retail marketing data, encompassing customer demographics, historical spending habits (on various product categories like wines, meat, etc.), and previous campaign responses, all necessary to predict the binary target variable, Response.

Organized and Detailed Summary of Milestones 1-3

Milestone 1: Data Selection and Exploratory Data Analysis (EDA)

The initial EDA confirmed the severity of the business problem and identified strong predictive signals.

- **Class Imbalance:** The analysis showed a severe class split: **85.0% Non-Response (0)** versus **15.0% Response (1)**. This confirmed that a high accuracy score is meaningless, necessitating the use of metrics focused on the minority class.

- **Demographic Signal (Education):** Customers with higher educational attainment (PhD, Master's) exhibited significantly higher acceptance rates (over 20%) compared to lower education

segments (below 5%), suggesting the current offering appeals to an affluent, established clientele.

- **Behavioral Signal (Recency):** The KDE plot demonstrated that converters are overwhelmingly clustered among **recently active customers** (low Recency values, peaking around 15 days). This established Recency as a key indicator for intervention.

- **Monetary Signal (Income & Spending):** The scatter plot showed that campaign acceptance is almost exclusively concentrated in the segment defined by **high income ($50k–$100k)** and **high overall spending** (high MntWines). This confirmed the model should focus on maximizing conversions within this high-value segment.

Milestone 2: Data Preparation and Feature Engineering

This milestone focused on creating a clean, numerical, and robust feature set for modeling.

1. Feature Drop Strategy

Features were systematically dropped to prevent data leakage and multicollinearity:

- **Data Leakage:** Past campaign outcomes (AcceptedCmp1 to AcceptedCmp5) were dropped as they directly predict the target (Response), leading to an artificially inflated and non-generalizable model.

- **Zero/Low Variance:** Constant columns (Z_CostContact, Z_Revenue) and near-zero variance columns (Complain) were dropped.

- **Non-Predictive:** Unique identifiers (ID) and redundant time variables (Dt_Customer) were excluded.

2. Imputation and Transformation

- **Missing Data:** The 24 missing values in the Income variable were handled using **Median Imputation** to avoid biasing the feature's heavily skewed distribution.

- **Feature Engineering:** Three powerful new features were engineered:
    - Total_Spent: Aggregating all monetary spending (a comprehensive CLV proxy).
    - Total_Purchases_Count: Aggregating total transactions (an overall engagement proxy).
    - Prop_Online_Purchases: Ratio of web/catalog purchases to total purchases (a channel preference indicator).

- **Encoding and Scaling:** Categorical features (Education, Marital_Status after grouping small segments) were converted to numerical **dummy variables**. Finally, all numerical features (Income, Total_Spent, channel counts) were normalized using **MinMaxScaler** to ensure fair contribution during model training.

Milestone 3: Model Building and Evaluation

1. Model Choice and Metrics

- **Model: Logistic Regression** was chosen due to its high **Interpretability** and ability to output a **Propensity Score** for ranking customers. The model used class_weight='balanced' to explicitly address the 85:15 data imbalance.

- **Primary Metric: F1-Score** for the positive class (Response=1), as it balances the need for high capture rate (Recall) against minimizing wasted spend (Precision).

2. Results and Performance

Metric	Score	Interpretation
AUC-ROC	**0.80**	Excellent discriminatory power (ability to rank positive cases higher than negative ones).
Recall (Class 1)	**0.65**	The model correctly identified 65% (65 out of 100) of the actual campaign responders.
Precision (Class 1)	**0.31**	When the model predicted a response, it was correct 31% of the time, leading to a 31% conversion rate among the targeted group.
F1-Score (Class 1)	**0.42**	The combined measure of positive performance.
Confusion Matrix	**TP=65, FN=35**	The model successfully captures the majority of converters while still having a manageable rate of False Positives (142 wasted targets).
Wasted Targets (FP)	**142**	Out of 572 non-responders, 142 were incorrectly flagged as high-propensity targets.

3. Feature Importance (Model Coefficients)

Rank	Feature	Coefficient	Impact
1	Prop_Online_Purchases	+3.66	**Strongest Positive Driver.** High ratio of online/catalog purchasing strongly increases conversion probability.
2	Total_Spent	+3.41	**Monetary Value.** High overall lifetime spending is the second strongest indicator.

3	NumWebVisitsMonth	+3.16	**Engagement.** Frequent browsing signals high interest.
15 (Worst)	NumStorePurchases	-1.20	**Strongest Negative Driver.** Customers who primarily shop in-store are highly unlikely to convert via this offer.
14	Education_Basic	-1.13	**Demographics.** Basic education is a major deterrent to acceptance.

Conclusion and Recommendations

1. What does the analysis/model building tell you?

The analysis confirms that the campaign is effectively targeting **high-value, highly engaged, and digitally-inclined customers.** The model successfully shifted the achievable conversion rate from a baseline of 15% (if targeting everyone) to **31%** (among the model-selected high-propensity group). The AUC of 0.80 proves the model is reliable for ranking customers by risk.

2. Is this model ready to be deployed?

Yes, the model is ready for a pilot deployment.

While the precision (31%) indicates that 69% of targeted high-propensity customers still do not convert (wasted spend), this is **more than double the current baseline conversion rate (15%)**. The high Recall (65%) ensures the company captures most of the potential revenue. The model output is a probability score, which is ideal: the marketing team can select a stricter threshold (e.g., target only customers with a score > 0.7) to further increase Precision and control costs for a smaller, highly profitable test group.

3. What are your recommendations?

1. **Prioritize Digital Spenders:** Stop blanket-targeting customers who rely on in-store purchases (NumStorePurchases is strongly negative). Focus campaign delivery and messaging entirely on online channels, particularly those demonstrating a high Prop_Online_Purchases ratio.

2. **Reward High Value:** Use Total_Spent as the primary filter. The campaign should be positioned as an exclusive offer for the company's established, high-monetary-value clientele.

3. **Target Engaged Browsers:** Leverage the positive coefficient on NumWebVisitsMonth. Target customers who are frequently browsing the website, even if they haven't purchased recently (low Recency), as this high engagement level indicates receptivity.

4. **Shift Budget:** Allocate the budget saved by excluding low-propensity segments toward improving the quality of the campaign for the targeted 31% Precision segment, potentially through more generous offers or higher-quality delivery methods.

4. Potential Challenges and Additional Opportunities

Area	Challenge	Opportunity
Data	**Feature Scaling:** The negative coefficient on NumWebPurchases despite the positive Prop_Online_Purchases suggests complex interactions or feature collinearity needs further investigation.	**New Features:** Engineer time-based features (e.g., time since enrollment, number of years as a customer) and more robust spending ratios.
Modeling	**Performance Ceiling:** Logistic Regression may have reached its performance limit (AUC 0.80).	**Alternative Models:** Test advanced models like **Random Forest** or **XGBoost** to potentially increase the AUC and F1-score.
Deployment	**Threshold Selection:** The 0.5 threshold used now is arbitrary.	**Cost-Benefit Analysis:** Perform a rigorous cost-benefit analysis at various thresholds to find the *optimal* score that maximizes ROI (profit), not just the F1-score.