Predicting Customer Response to Marketing Campaigns Using Machine Learning

By Bradley Bryant

Executive Summary

Marketing campaigns are powerful tools to boost customer engagement and revenue, but when poorly targeted, they waste budget and lower ROI. This project applies predictive analytics to identify customers most likely to respond to future marketing campaigns, using machine learning on the Customer Personality dataset. Through Logistic Regression and Random Forest models, supported by comprehensive data cleaning and engineered features, we built a predictive pipeline that empowers more precise and effective marketing strategies.

Introduction and Problem Statement

Effective marketing is no longer about mass messaging; it's about precision and personalization. Blanket campaigns may reach a wide audience but often suffer from low engagement and high customer attrition. Businesses need analytical tools that can help them direct marketing efforts towards customers that will be most likely to respond, increasing campaign ROI and reducing wasted efforts. This project explores the application of machine learning to predict customer responsiveness to direct marketing campaigns. Our core question is: Can we use customer data to predict which individuals are likely to respond to a new campaign? The dataset used comes from Kaggle and includes records for 2,240 customers of a Portuguese retailer. The data includes demographic details, purchasing habits, campaign engagement history, and interaction frequency. Our target variable is Response, which indicates whether the customer accepted the most recent campaign (1) or not (0). By developing a binary classification model using this variable, our goal is to create a tool that enables smarter, data-informed campaign targeting.

Data Overview

The dataset, *sourced from Kaggle's Customer Personality Analysis (Martins, 2015).* includes 29 variables, categorized as:

People

- ID: Customer's unique identifier
- Year_Birth: Customer's birth year
- Education: Customer's education level
- Marital_Status: Customer's marital status
- Income: Customer's yearly household income
- Kidhome: Number of children in customer's household
- Teenhome: Number of teenagers in customer's household
- Dt_Customer: Date of customer's enrollment with the company
- Recency: Number of days since customer's last purchase
- Complain: 1 if the customer complained in the last 2 years, 0 otherwise

Products

- MntWines: Amount spent on wine in last 2 years
- MntFruits: Amount spent on fruits in last 2 years
- MntMeatProducts: Amount spent on meat in last 2 years
- MntFishProducts: Amount spent on fish in last 2 years
- MntSweetProducts: Amount spent on sweets in last 2 years
- MntGoldProds: Amount spent on gold in last 2 years

Promotion

- NumDealsPurchases: Number of purchases made with a discount
- AcceptedCmp1: 1 if customer accepted the offer in the 1st campaign, 0 otherwise
- AcceptedCmp2: 1 if customer accepted the offer in the 2nd campaign, 0 otherwise
- AcceptedCmp3: 1 if customer accepted the offer in the 3rd campaign, 0 otherwise

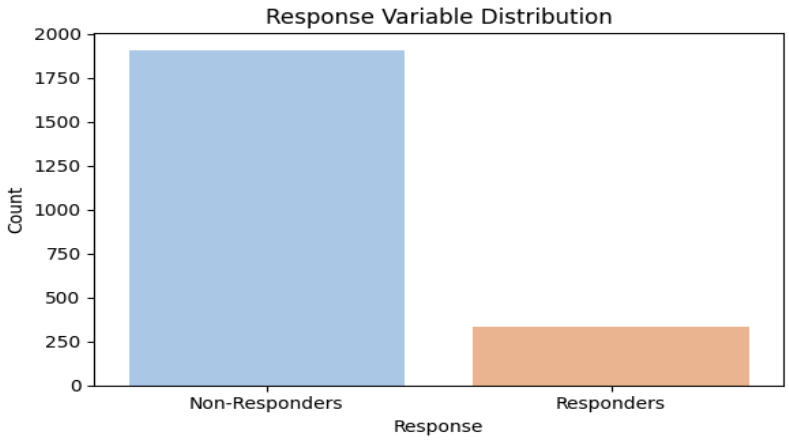
- AcceptedCmp4: 1 if customer accepted the offer in the 4th campaign, 0 otherwise
- AcceptedCmp5: 1 if customer accepted the offer in the 5th campaign, 0 otherwise
- Response: 1 if customer accepted the offer in the last campaign, 0 otherwise

Place

- NumWebPurchases: Number of purchases made through the company's website
- NumCatalogPurchases: Number of purchases made using a catalogue
- NumStorePurchases: Number of purchases made directly in stores
- NumWebVisitsMonth: Number of visits to company's website in the last month

Data Cleaning and Preprocessing

Our first step was handling the missing values in the Income column using median imputation, which is robust to outliers and skewed distributions. We then dropped constant or non-informative columns: Z_CostContact, Z_Revenue, and the unique identifier ID, which had no predictive value.
The target variable, Response, exhibited a significant imbalance—only ~15% of customers responded to the campaign, while the remaining 85% did not. To illustrate this, we generated a bar chart displaying the counts of responders and non-responders.



This imbalance posed a risk of bias in model predictions and was later addressed in the modeling phase through class weighting and threshold tuning techniques.

To reduce skewness in income and stabilize variance, we applied a log transformation using np.log1p() and replaced the original Income column with the transformed values.

We then addressed inconsistencies in the categorical variables. For Marital_Status, values such as “Divorced,” “Widow,” “Alone,” “Absurd,” and “YOLO” were consolidated into a single “Single” category to reduce fragmentation and improve interpretability. For Education, we regrouped similar levels to generalize tiers: “2n Cycle” was reclassified as “Master,” and “Graduation” as “Bachelor.” After recategorizing, we applied one-hot encoding to both variables using pd.get_dummies() with drop_first=True. This created binary features for Master, PhD, and Basic under Education, with Bachelor as the reference category, and Single and Together under Marital_Status, with Married as the reference. This transformation allowed the models to capture categorical patterns while avoiding multicollinearity.

We also identified and removed age outliers, customers whose calculated age exceeded 100 years, ensuring the dataset reflected plausible demographics.

All numerical variables were standardized using Z-score normalization with StandardScaler to ensure comparability across features and support effective model training.

Feature Engineering

To enhance model performance and capture behavioral patterns, we created the following features:

- Age: Derived from Year_Birth
- Customer_Tenure_Years: Time since the customer joined (previously created during preprocessing)
- TotalSpent: Sum of spending across all product categories (MntWines, MntMeatProducts, etc.)

- TotalPurchases: Sum of NumWebPurchases, NumCatalogPurchases, and NumStorePurchases
- TotalAcceptedCampaigns: Sum of all five AcceptedCmp variables
- Children: Sum of Kidhome and Teenhome

These features helped quantify loyalty, customer value, and engagement behavior.

Additionally, we created binned features such as SpendingGroup, LoyaltyGroup, RecencyGroup, and CampaignGroup for exploratory data analysis (EDA). These variables were used to visualize response patterns but were not included in the final modeling pipeline.

Modeling Approach

We developed and evaluated two classification models to predict campaign response:

- Logistic Regression: A linear and interpretable baseline model
- Random Forest Classifier: A nonlinear, ensemble-based model known for robustness and automatic feature ranking

To address the class imbalance in the target variable (~15% positive class), we used class_weight='balanced' in both models. This adjustment ensured that the minority class (responders) received appropriate weight during training.

We applied Sequential Forward Feature Selection to the Logistic Regression model to reduce dimensionality and improve generalization. The top 15 features were selected based on AUC performance using 5-fold cross-validation. For the Random Forest, we used all available features, since the model handles high-dimensional data effectively and performs internal feature selection.

The dataset was split into 80% training and 20% testing sets using stratified sampling to preserve the original response distribution. All numerical variables were standardized

using Z-score normalization prior to training to ensure comparability across features, especially for Logistic Regression.

Model Evaluation and Threshold Tuning

We evaluated model performance using several key metrics: Area Under the ROC Curve (AUC), F1 Score, Precision, and Recall. Initially, both models were assessed using the default decision threshold of 0.50. However, to better balance precision and recall, we performed threshold tuning on the Random Forest model using a custom loop to identify the threshold that maximized F1 Score.

The final comparison across models is shown below:

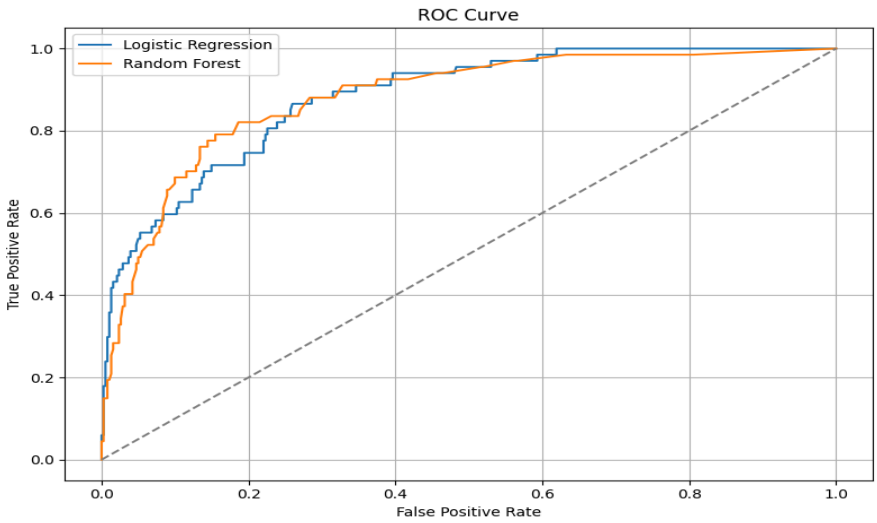
Model	AUC	F1 Score	Precision	Recall	Threshold
Logistic Regression (0.5)	0.881	0.55	0.453	0.716	0.50
Random Forest (0.5)	0.881	0.413	0.760	0.284	0.50
Random Forest (0.26)	0.881	0.600	0.542	0.672	0.26

Lowering the threshold from 0.50 to 0.26 significantly increased recall from 0.284 to 0.672, while maintaining a respectable precision of 0.542. This shift led to a balanced F1 score of 0.600, representing the optimal trade-off between false positives and false negatives.

In the context of marketing, this trade-off is highly valuable: it is often more acceptable to contact a few uninterested customers than to miss a high-value potential responder. A threshold of 0.26 therefore aligns with business goals by maximizing campaign reach among likely converters without sacrificing too much targeting precision.

In addition to precision, recall, and F1 Score, we also evaluated sensitivity and specificity to further understand the model's ability to correctly classify both responders and non-responders. Sensitivity, also known as the true positive rate, reflects how well the model identifies actual responders, while specificity measures how effectively it avoids false positives. The threshold-tuned Random Forest model achieved a balanced profile across these metrics, indicating its suitability for campaign

targeting where capturing true responders is essential, but excessive misclassification of non-responders must also be avoided. For the Random Forest model tuned at threshold = 0.26, sensitivity was 0.672 and specificity was 0.900, confirming its strong ability to detect likely responders while minimizing wasted outreach on low-probability customers.



Feature Importance and Interpretation

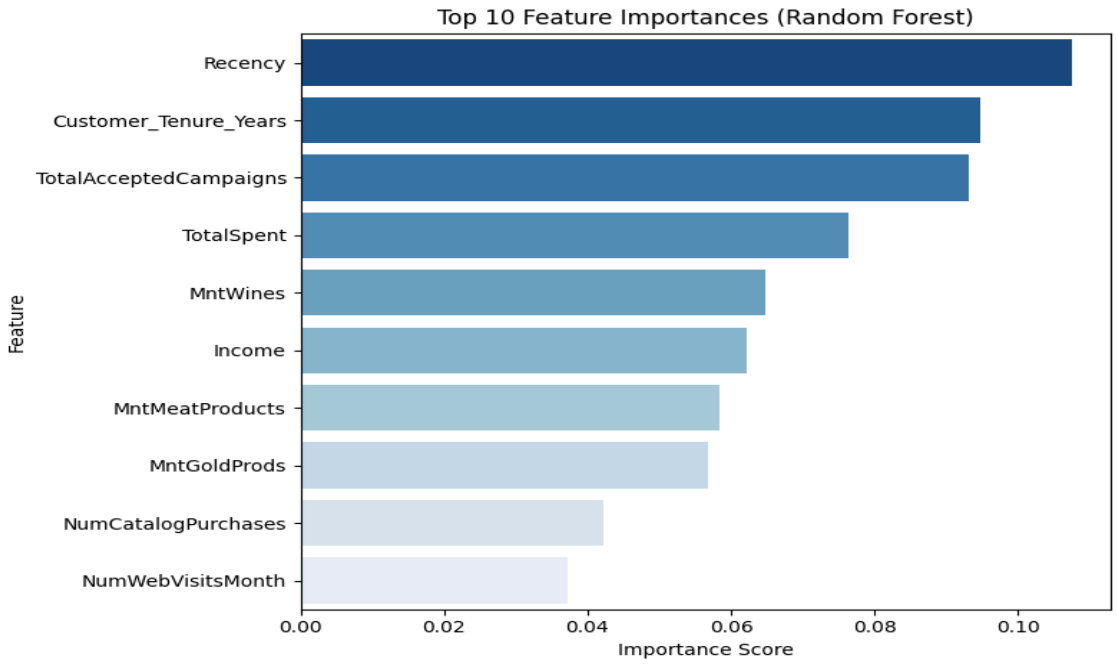
Understanding which features drive model predictions is critical for translating machine learning output into actionable marketing strategies. Both the Random Forest and Logistic Regression models offered interpretable insights into customer behaviors that most influenced campaign response.

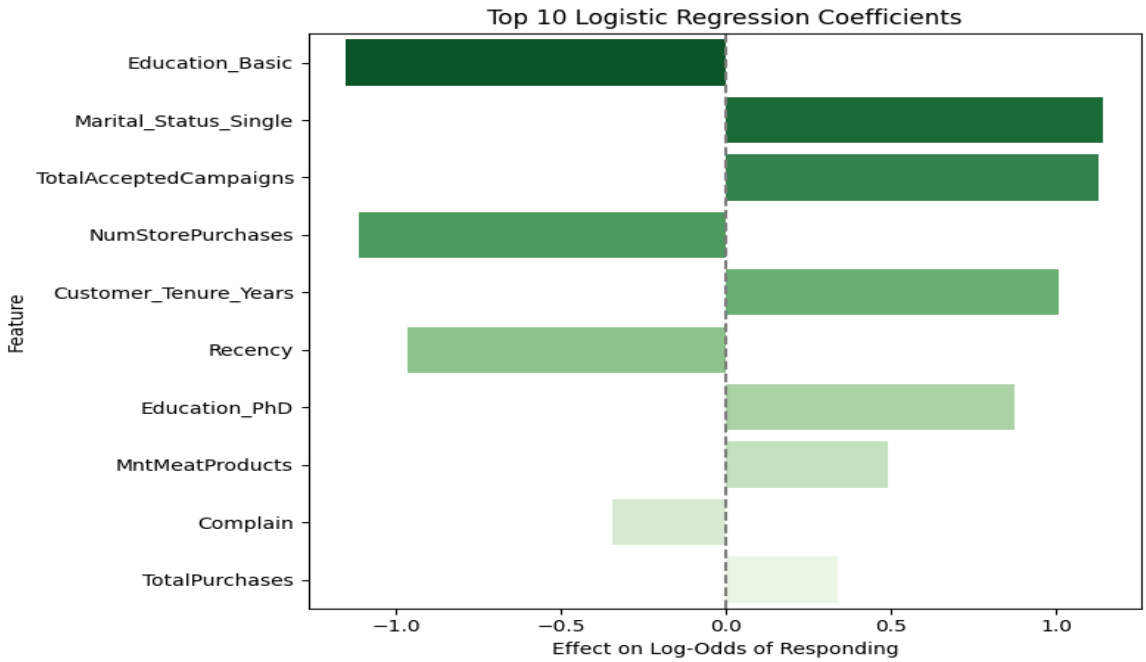
Random Forest ranked features by their contribution to reducing classification error, while Logistic Regression provided directional insights—revealing not only which traits were influential, but also whether they increased or decreased the likelihood of response. This complementarity allowed for a more nuanced understanding of the predictive drivers.

Variables such as Recency, Customer_Tenure_Years, and TotalAcceptedCampaigns were top predictors in both models, reinforcing their importance across modeling approaches.

The top 10 predictors from each model are summarized below:

Random Forest - Top Features	**Logistic Regression - Top Coefficients**
Recency	Education_Basic (-)
Customer_Tenure_Years	Marital_Status_Single (+)
TotalAcceptedCampaigns	TotalAcceptedCampaigns (+)
TotalSpent	NumStorePurchases (-)
MntWines	Customer_Tenure_Years (+)
Income (log-transformed)	Recency (-)
MntMeatProducts	Education_PhD (+)
MntGoldProds	MntMeatProducts (+)
NumCatalogPurchases	Complain (-)
NumWebVisitsMonth	TotalPurchases (-)





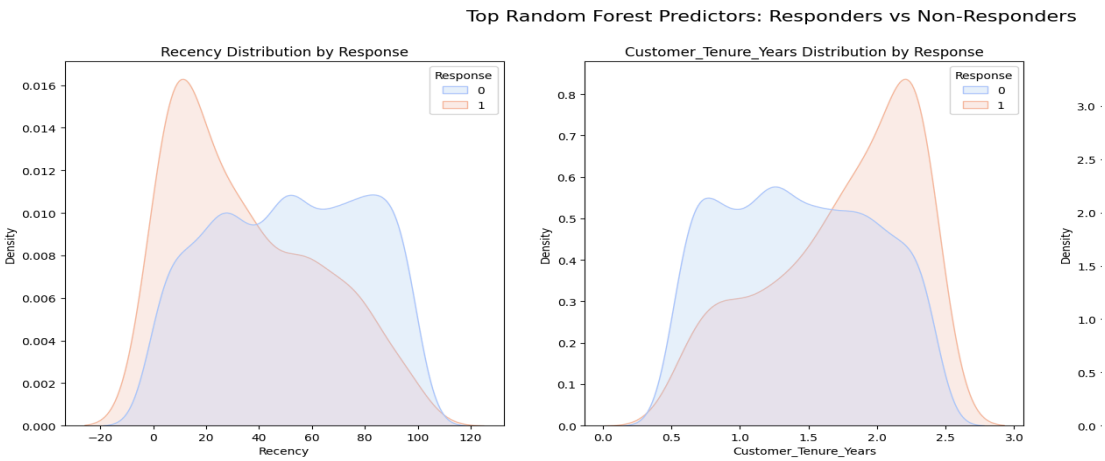
Key Takeaways:

- Recency and Customer_Tenure_Years were consistently strong predictors across both models, emphasizing that loyal and recently active customers are the most responsive to campaigns.

- TotalAcceptedCampaigns emerged as a robust signal of future engagement, validating that prior campaign success is a key indicator for targeting.

- Logistic Regression adds value by showing directionality: for instance, customers with basic education or high in-store purchases were less likely to respond, while single customers and PhD holders showed increased responsiveness.

To visually reinforce these findings, kernel density plots (KDEs) were created for top features. Notably, Recency and Customer_Tenure_Years displayed clear separation between responders and non-responders, underscoring their predictive power in real-world campaign targeting.



Insights from Data Exploration

Exploratory visualizations revealed two dominant behavioral patterns among responders: Customer Loyalty and High Spending Behavior.

Insight 1: Customer Loyalty Drives Campaign Response

Three key indicators of customer loyalty tenure, recency, and total accepted campaigns were all strongly associated with higher response rates.

The tenure bar plot and TenureGroup heat map showed a clear upward trend: customers who had been with the company for over two years had significantly higher response rates, particularly those who also belonged to the top spending quartile.

Recency visualizations confirmed that recent interaction is a powerful predictor of responsiveness. Customers in the 0–10 day recency group had response rates exceeding 30%, with steep declines observed as recency increased beyond 90 days.

The Total Accepted Campaigns plot showed a consistent increase in response likelihood with each additional campaign accepted. Customers who had accepted four to five previous campaigns had response rates approaching 90%, confirming that past engagement is a strong signal of future responsiveness.

A joint bar plot of RecencyGroup and CampaignGroup reinforced this pattern, showing that customers with both recent interactions and multiple prior campaign acceptances were almost certain to respond again, with response rates approaching 100%.

Together, these patterns highlight a clear business opportunity: customers who are loyal, recently active, and have historically responded to offers should be prioritized in future campaigns.

Insight 2: High-Spending Customers Are More Likely to Respond

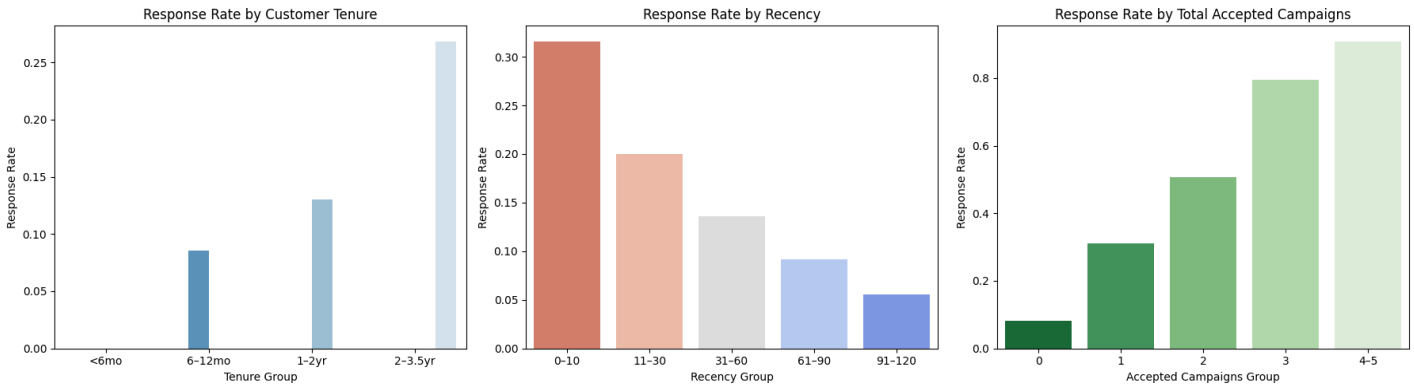
Spending behavior was another key predictor of engagement.

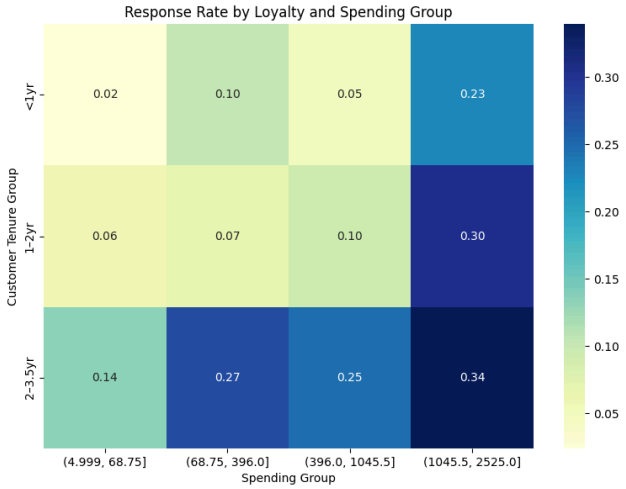
The heatmap of SpendingGroup vs. LoyaltyGroup revealed that customers in the top quartile of spending and with over two years of tenure had the highest overall response rates, exceeding 30% in several segments.

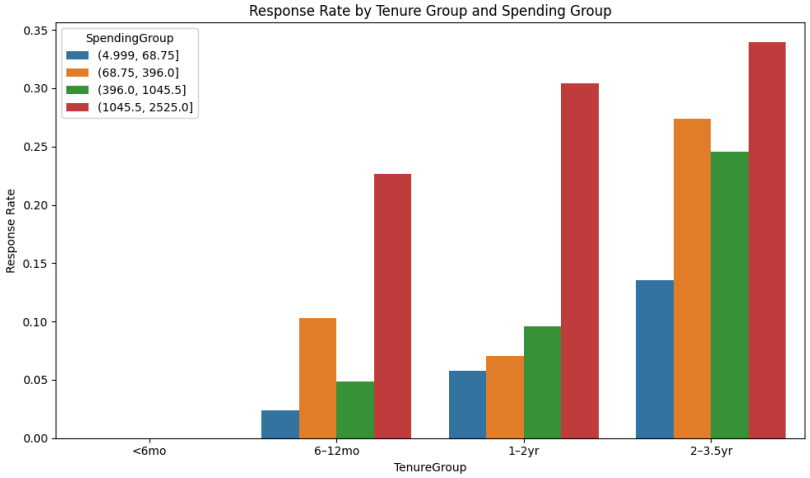
Similarly, the bar plot of TenureGroup vs. SpendingGroup showed that even customers with shorter tenure (6–12 months) but high spending levels responded more than low spenders with longer tenure.

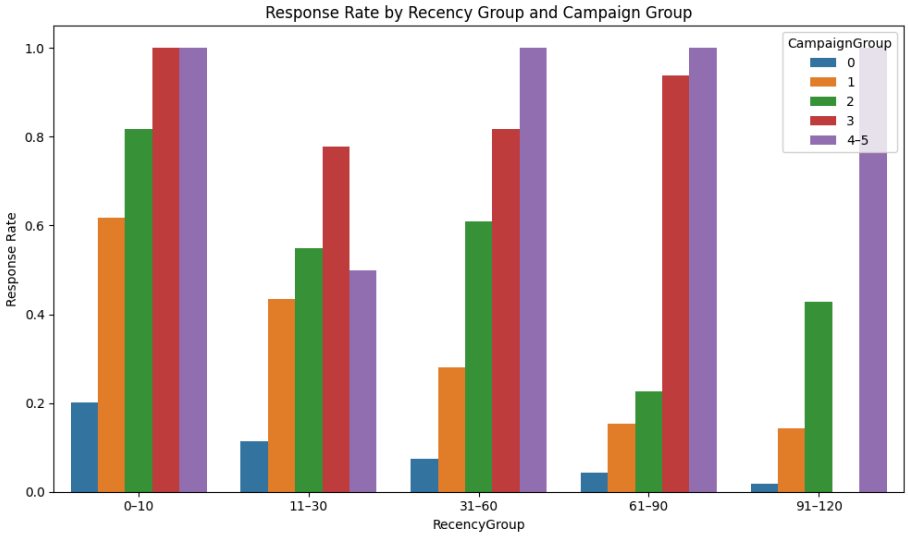
These visualizations confirm that financial value and loyalty reinforce one another. While long-term, high-spending customers form the most responsive core, even newer customers with early high-value behavior show promising potential for future engagement.

Together, these insights validate the model's emphasis on features like Customer_Tenure_Years, Recency, TotalAcceptedCampaigns, and TotalSpent. From a strategic marketing perspective, they support a dual-focus: retain and reward loyal customers, while also nurturing new high-potential customers early in their lifecycle.









Customer Targeting and Business Application

Using the final Random Forest model with a tuned threshold of 0.26, we scored all customers by their predicted probability of responding to future marketing campaigns. The top-scoring individuals shared common traits aligned with the model's key predictors—recent engagement, long tenure (1.2–2.2 years), and responsiveness to previous campaigns (accepting two to four prior offers).

Their purchasing behavior further emphasized their value: frequent transactions (14–27 total purchases), strong activity across both store and web channels, and notable

spending in meat and wine categories. Additionally, many had one or more dependents, highlighting a family-oriented consumer profile. These customers represent ideal candidates for future outreach, reflecting personas such as multi-channel shoppers and lifestyle-driven buyers.

To act on these insights, the model can be integrated into a company's CRM or marketing platform to enable real-time scoring and campaign automation. High-probability responders can be prioritized across email, SMS, and direct mail channels, with tailored offers based on past behavior. The model also supports dynamic segmentation, lookalike audience creation for acquisition, and A/B testing across score tiers. Retraining the model regularly with new data ensures it adapts to evolving customer behavior.

By embedding this predictive system into marketing operations, businesses can reduce wasted spend, increase conversion rates, and scale smarter, more personalized campaigns.

Conclusion and Strategic Recommendations

This project addressed a central challenge in modern marketing: how to effectively identify and prioritize customers most likely to respond to a campaign. By implementing Random Forest and Logistic Regression within a structured machine learning pipeline including, data preprocessing, feature engineering, model training, and evaluation, we developed a predictive system that transitions marketing efforts from reactive to proactive.

Both models produced statistically sound and interpretable results. Key features such as recency, customer tenure, total spending, and prior campaign engagement consistently emerged as the most influential predictors of response. These findings were reinforced through exploratory visualizations, lending both behavioral and statistical credibility to the model outputs.

A notable strength of the Random Forest model was its ability to capture complex feature interactions while maintaining strong performance despite class imbalance. Logistic Regression, while simpler, provided clear directional insights and served as a reliable baseline. Together, they offered a balanced mix of performance and interpretability.

There are, however, a few limitations to acknowledge. The models rely on structured historical data and may not generalize well without regular retraining. While class weighting mitigated imbalance, alternative approaches such as SMOTE or cost-sensitive learning could further enhance recall.

Future improvements could involve more advanced models like XGBoost, hyperparameter tuning, and the integration of unstructured behavioral data sources to enrich segmentation.

In summary, this project demonstrates how predictive analytics can guide data-driven marketing. With further refinement, the model can help businesses improve campaign ROI and strengthen customer relationships at scale

References

Martins, R. (2015). *Customer Personality Analysis* [Data set]. Kaggle. https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis