Modeling Customer Responsiveness to Marketing Campaigns Using Machine Learning

Anisha Choudhury and Tu Tong
Statistical and Machine Learning, Data Analytics Department, Dickinson College

Introduction

Motivation:

- Traditional marketing strategies are becoming less effective as customers grow more informed and selective.
- Retailers are not only able to access purchasing history but also store customers' socio-demographic data and utilize it to enhance their strategies and customer satisfaction.
- To stay competitive, businesses must understand individual preferences and behavioral patterns and deliver personalized campaigns.

Objective:

- Extract important characteristics used for predicting response.
- Analyze the relationship between these characteristics and the likelihood of responding to marketing campaigns
- Generate interpretable insights that can guide targeted marketing and optimize campaign effectiveness.

Data and Features

Kaggle dataset: 29 columns (initial 28 features columns and 1 response column) and 2240 rows.

Some of the important features in the dataset:

People

- 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
- Recency: Number of days since customer's last purchase

Products

- MntWines/Fruits/Meat/Fish/Sweet/Gold: Amount spent on wine, fruits, meat, fish, sweet, gold in last 2 years.

Promotion

- NumDealsPurchases: Number of purchases made with a discount
- AcceptedCmp1/2/3/4/5: 1 if customer accepted the offer in the nth campaigns, 0 otherwise.
- Response: 1 if customer accepted the offer in the last campaign, 0 otherwise

Data Preparation steps:

1. Data Cleaning

- Remove rows with missing values and construct one-hot encoding for categorical feature (marital_status) and label encoding for ordinal feature (education).
- Normalize income to account for right-skewed distribution.
- Drop numerical and categorical features with significantly lower correlation.

2. Data Preprocessing

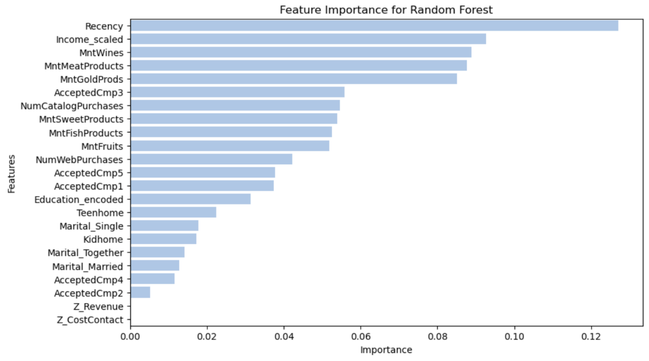
- To address class imbalance in our binary response variable (~1,400 non-responders vs. ~300 responders), we applied oversampling on the training set to create a balanced dataset.
- All models were trained using the balanced dataset to ensure unbiased and consistent comparisons.
- A stratified 80/20 train-test split was used to evaluate model performance and generalizability.

Methodology

Feature Selection

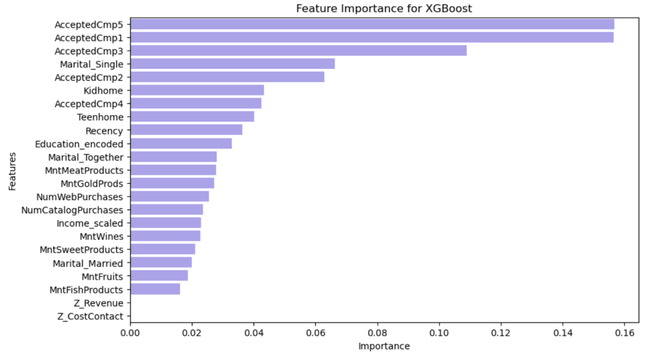
1. Random Forest

- Random Forest was used primarily for feature selection. This ensemble method builds multiple decision trees on bootstrapped samples and averages their results, reducing overfitting and improving stability. Its inherent ability to rank features by importance made it a valuable tool for narrowing down predictive variables.
- Hyperparameters (n_estimators, max_depth, min_samples_split, min_samples_leaf, and max_features) were tuned using grid search with cross-validation to optimize model performance. The top features identified here were retained for further modeling.



2. XGBoost

- Extreme Gradient Boosting was employed as the second feature selection method and benchmark model.
- Known for its scalability and predictive power, XGBoost iteratively improves weak learners to minimize loss.
- Hyperparameters for tuning: n_estimators, max_depth, learning_rate, subsample, colsample_bytree, gamma, reg_alpha, and reg_lambda.



Model Performance Evaluation

MODEL	METRICS			
	Accuracy	Precision	Recall	F1 Score
Random Forest	0.8828	0.6829	0.4179	0.5185
XGBoost	0.8671	0.5645	0.5224	0.5426

While Random Forest delivers better accuracy and precision, XGBoost achieves slightly higher recall and F1 scores. Both models demonstrate strong overall performance after tuning, with accuracy around 0.8, but there is still room for improvement.

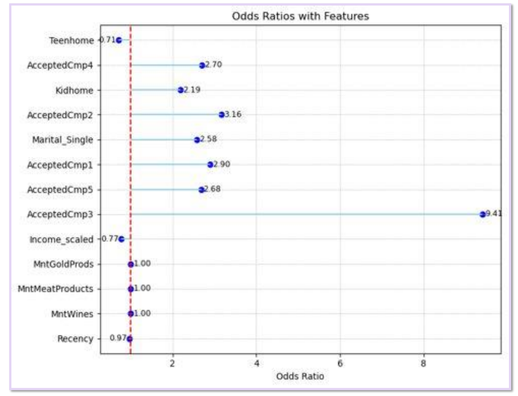
Interpretation model

Linear Regression Model

- Features that were consistently ranked highly by both Random Forest and XGBoost were selected for the final logistic regression model to ensure robustness.
- Logistic regression was applied to interpret the relationship between customer characteristics and their likelihood of responding to the most recent campaign. It is well-suited for binary classification and offers strong interpretability through odds ratios, which quantify both the magnitude and direction of each feature's effect.
- Post-training, model coefficients were exponentiated to obtain odds ratios. These values helped interpret how each predictor influenced the response. Features with odds ratios >1 were positively associated with acceptance, while those <1 showed a negative association.

Results

The following is the Odds Ratio result:



Several variables showed positive associations i.e more likely to accept future campaigns:

- Customers who had previously accepted offers were significantly more likely to respond again.
- Specifically, accepting Campaign 1, 2, 4, or 5 increased the odds of acceptance by roughly 3x, 3x, 2.5x, and 2.5x, respectively.
- Notably, customers who accepted Campaign 3 were over 9x more likely to accept the current offer, underscoring a strong effect of past engagement.
- Each additional child in the household more than doubled the odds of acceptance, increasing them to approximately 2x.
- Customers who were single had around 2.5x higher odds of accepting compared to those who were married or in other relationship categories.
- Spending behaviors also had minor but directionally consistent effects.
- Each additional teenager reduced the odds of acceptance by approximately 29%.
- Similarly, higher income levels were associated with a decreased likelihood of response.

Conclusion

Using machine learning models for feature selection, we identified 13 key predictors influencing customer response to a marketing campaign. Our final logistic regression model achieved ~72% accuracy, and odds ratio interpretation provided clear insights into how specific customer characteristics impact campaign engagement.